Statistical Science
2003, Vol. 18, No. 1, 71–103

Multiple Hypothesis Testing in Microarray Experiments

Sandrine Dudoit, Juliet Popper Shaffer and Jennifer C. Boldrick

Abstract. DNA microarrays are part of a new and promising class of biotechnologies that allow the monitoring of expression levels in cells for thousands of genes simultaneously. An important and common question in DNA microarray experiments is the identification of differentially expressed genes, that is, 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 DNA microarray experiments and compares the procedures on microarray and simulated data sets.

Key words and phrases: Multiple hypothesis testing, adjusted p-value, family-wise Type I error rate, false discovery rate, permutation, DNA microarray.

1. INTRODUCTION

The burgeoning field of genomics has revived interest in multiple testing procedures by raising new methodological and computational challenges. For example, DNA microarray experiments generate large multiplicity problems in which thousands of hypotheses are tested simultaneously. DNA microarrays are high-throughput biological assays that can measure DNA or RNA abundance 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 DNA microarray experiments is the identification of differentially expressed genes, that is, 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

Sandrine Dudoit is Assistant Professor, Division of Biostatistics, School of Public Health, University of California, Berkeley, California 94720-7360 (e-mail: sandrine@stat.berkeley.edu). Juliet Popper Shaffer is Senior Lecturer Emerita, Department of Statistics, University of California, Berkeley, California 94720-3860 (e-mail: shaffer@stat.berkeley.edu). Jennifer C. Boldrick is a graduate student, Department of Microbiology and Immunology, Stanford University, Stanford, California 94305-5124 (e-mail: boldrick@stanford.edu).

is not, 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 no longer corresponds to a significant finding, as it is very likely that such a small p-value will occur by chance under the null hypothesis when considering a large enough set of genes. Special problems that arise from the multiplicity aspect include defining an appropriate Type I error rate and devising powerful multiple testing procedures that control this error rate and account for the joint distribution of the test statistics. A number of recent articles have addressed the question of multiple testing in DNA microarray experiments. However, the proposed solutions have not always been cast in the standard statistical framework (Dudoit et al., 2002; Efron et al., 2000; Golub et al., 1999; Kerr, Martin and Churchill, 2000; Manduchi et al., 2000; Pollard and van der Laan, 2003; Tusher, Tibshirani and Chu, 2001; Westfall, Zaykin and Young, 2001).

The present article discusses different approaches to multiple hypothesis testing in the context of DNA microarray experiments and compares the procedures on microarray and simulated data sets. Section 2 reviews basic notions and procedures for multiple testing, and discusses the recent proposals of Golub et al. (1999) and Tusher, Tibshirani and Chu (2001) within this framework. The microarray data sets 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 5summarizes our findings and outlines open questions. Although the focus is on the identification of differentially expressed genes in DNA 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 DNA Microarray Experiments

Consider a DNA microarray experiment which produces expression data on m genes (i.e., variables or features) for n mRNA samples (i.e., observations), 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 measure 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 discussed in Lipshutz et al. (1999)] or relative with respect to the expression levels of a suitably defined common reference sample [e.g., Stanford two-color spotted cDNA microarrays discussed in Brown and Botstein (1999)]. Note that the expression measures 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 et al., 2001, 2002). The gene expression data are conventionally stored in an $m \times n$ matrix $X = (x_{ji})$, with rows corresponding to genes and columns corresponding to individual mRNA samples. In a typical experiment, the total number n of samples is anywhere between around 10 and a few hundreds, and the number m of genes is several thousands. The gene expression measures, x, are generally continuous variables, while the responses or covariates, y, can be either polytomous or continuous, and possibly censored, as described above.

The pairs $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$, formed by the expression profiles $\mathbf{x}_i$ and responses or covariates y_i, are viewed as a random sample from a population of interest. The population and sampling mechanism will depend on the particular application (e.g., designed factorial experiment in yeast, retrospective study of human tumor gene expression). Let X_j and Y denote, respectively, the random variables that correspond to the expression measure for gene j, $j = 1, \ldots, m$, and the response or covariate. The goal is to use the sample data $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$ to make inference about the population of interest, specifically, test hypotheses concerning the joint distribution of the expression measures $\mathbf{X} = (X_1, \ldots, X_m)$ and response or covariate Y.

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 the expression measure X_j and the response or covariate Y. (In some cases, more specific null hypotheses may be

of interest, for example, the null hypothesis of equal mean expression levels in two populations of cells as opposed to identical distributions.) A standard approach to the multiple testing 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, Martin and Churchill, 2000; Manduchi et al., 2000; Pollard and van der Laan, 2003; Tusher, Tibshirani and Chu, 2001; Westfall, Zaykin and Young, 2001).

The univariate problem 1 has been studied extensively in the statistical literature (Lehmann, 1986). 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-statistic or a Mann–Whitney statistic; for polytomous responses, one might use an F-statistic; and for survival data one might rely on the score statistic for the Cox proportional hazards model. We will not 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 which is a function of X_j and Y. The lower case 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, DNA microarray experiments present a new and challenging area of application for multiple testing procedures because of the sheer number of tests. In the remainder of this section, we review basic notions and approaches to multiple testing and discuss recent proposals for dealing with the multiplicity problem in microarray experiments.

2.2 Type I and Type II 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 Table 1 (Benjamini and 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. Ideally, 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*. A standard approach in the univariate setting is to prespecify an acceptable level α for the Type I error rate and seek tests which minimize the Type II error rate, that is, maximize *power*, within the class of tests with Type I error rate at most α.

TABLE 1

Number of	Number not rejected	Number 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, that is, 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 $\Pr(|T_1| \geq c_\alpha \mid H_1) \leq \alpha$ 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).

- *The per-comparison error rate* (PCER) is defined as the expected value of the number of Type I errors divided by the number of hypotheses, that is, $\text{PCER} = E(V)/m$.
- *The per-family error rate* (PFER) is defined as the expected number of Type I errors, that is, $\text{PFER} = E(V)$.
- *The family-wise error rate* (FWER) is defined as the probability of at least one Type I error, that is, $\text{FWER} = \Pr(V \geq 1)$.
- The *false discovery rate* (FDR) of Benjamini and Hochberg (1995) is the expected proportion of Type I errors among the rejected hypotheses, that is, $\text{FDR} = E(Q)$, where, by definition, $Q = V/R$ if $R > 0$ and 0 if $R = 0$.

A multiple testing procedure is said to control a particular Type I error rate at *level* α, if this error rate is less than or equal to α when the given procedure is applied to produce a list of R rejected hypotheses. For instance, the FWER is controlled at level α by

a particular multiple testing procedure if FWER $\leq \alpha$ (similarly, for the other definitions of Type I error rates).

Strong versus weak control. It is important to note that the error rates above are defined under the true and typically unknown data generating distribution for $\mathbf{X} = (X_1, \ldots, X_m)$ and Y. In particular, they depend upon which specific subset $\Lambda_0 \subseteq \{1, \ldots, m\}$ of null hypotheses is true for this distribution. That is, the family-wise error rate is FWER $= \Pr(V \geq 1 \mid \bigcap_{j \in \Lambda_0} \mathrm{H}_j) = \Pr(\text{Reject at least one } \mathrm{H}_j,\ j \in \Lambda_0 \mid \bigcap_{j \in \Lambda_0} \mathrm{H}_j)$, where $\bigcap_{j \in \Lambda_0} \mathrm{H}_j$ refers to the subset of true null hypotheses for the data generating joint distribution. A fundamental, yet often ignored distinction, is that between strong and weak control of a Type I error rate (Westfall and Young, 1993, page 10). *Strong control* refers to control of the Type I error rate under any combination of true and false null hypotheses, i.e., for any subset $\Lambda_0 \subseteq \{1, \ldots, m\}$ of true null hypotheses. In contrast, *weak control* refers to control of the Type I error rate only when all the null hypotheses are true, i.e., under a null distribution satisfying the *complete null hypothesis* $\mathrm{H}_0^C = \bigcap_{j=1}^m \mathrm{H}_j$ with $m_0 = m$. In general, the complete null hypothesis H_0^C is not realistic and weak control is unsatisfactory. In reality, some null hypotheses Λ_0 may be true and others false, but the subset Λ_0 is unknown. Strong control ensures that the Type I error rate is controlled under the true and unknown data generating distribution. 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. Note that the concept of strong and weak control applies to each of the Type I error rates defined above, PCER, PFER, FWER and FDR. The reader is referred to Pollard and van der Laan (2003) for a discussion of multivariate null distributions and proposals for specifying such joint distributions based on projections of the data generating distribution or of the joint distribution of the test statistics on submodels satisfying the null hypotheses. In the remainder of this article, unless specified otherwise, probabilities and expectations are computed for the combination of true and false null hypotheses corresponding to the true data generating distribution, that is, under the composite null hypothesis $\bigcap_{j \in \Lambda_0} \mathrm{H}_j$ corresponding to the data generating distribution, where $\Lambda_0 \subseteq \{1, \ldots, m\}$ is 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 (1) the probability of rejecting at least one false null hypothesis, $\Pr(S \geq 1) = \Pr(T \leq m_1 - 1)$, (2) the average probability of rejecting the false null hypotheses, $E(S)/m_1$, or *average power*, and (3) 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 $E(S/R \mid R > 0) \Pr(R > 0) = \Pr(R > 0) - \mathrm{FDR}$; when $m = m_1$, this is the any-pair power $\Pr(S \geq 1)$. One should note again that probabilities depend upon which particular subset $\Lambda_0 \subseteq \{1, \ldots, m\}$ of null hypotheses is true.

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 that control the PFER are generally more *conservative*, that is, lead to fewer rejections, than those that control either the FWER or the PCER, and procedures that control the FWER are more conservative than those that control 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 also involves the *joint* distribution of the test statistics T_j:

$$\mathrm{PCER} = \frac{\alpha_1 + \cdots + \alpha_m}{m} \leq \max(\alpha_1, \ldots, \alpha_m) \leq \mathrm{FWER} \leq \mathrm{PFER} = \alpha_1 + \cdots + \alpha_m.$$

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 (Benjamini and Hochberg, 1995). The classical approach to multiple testing calls for strong control of the FWER (cf. Bonferroni procedure). The recent proposal of Benjamini and Hochberg (1995) controls the FWER in the weak sense (since FDR $=$ FWER under the complete null) and can be less conservative than FWER

otherwise. Procedures that control the PCER are generally less conservative than those that control 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 Gaussian random 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, which equals 1 if the condition in parentheses is true and 0 otherwise. Assume without loss of generality that the m_0 true null hypotheses are $\mathrm{H}_1, \ldots, \mathrm{H}_{m_0}$, that is, $\Lambda_0 = \{1, \ldots, m_0\}$. Then $V = \sum_{j=1}^{m_0} R_j$ and $R = \sum_{j=1}^{m} R_j$. Analytical formulae for the Type I error rates can easily be derived as $\mathrm{PFER} = \sum_{j=1}^{m_0} \gamma_j$, $\mathrm{PCER} = \sum_{j=1}^{m_0} \gamma_j/m$, $\mathrm{FWER} = 1 - \prod_{j=1}^{m_0}(1 - \gamma_j)$ and

$$\mathrm{FDR} = \sum_{r_1=0}^{1} \cdots \sum_{r_m=0}^{1} \frac{\sum_{j=1}^{m_0} r_j}{\sum_{j=1}^{m} r_j} \prod_{j=1}^{m} \gamma_j^{r_j} (1 - \gamma_j)^{1-r_j}$$

with the FDR convention that $0/0 = 0$ and $\gamma_j = E(R_j) = \Pr(R_j = 1) = 1 - \Phi(z_{\alpha/2} - \mu_j\sqrt{n}) + \Phi(-z_{\alpha/2} - \mu_j\sqrt{n})$ 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 $\mathrm{PFER} = m_0\alpha$, $\mathrm{PCER} = m_0\alpha/m$, $\mathrm{FWER} = 1 - (1 - \alpha)^{m_0}$ and

$$\mathrm{FDR} = \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \binom{m_0}{v} \alpha^v (1-\alpha)^{m_0 - v} \times \binom{m_1}{s} \beta^s (1-\beta)^{m_1 - s},$$

where $\beta = 1 - \Phi(z_{\alpha/2} - d) + \Phi(-z_{\alpha/2} - d)$. Note that unlike the PCER, PFER or FWER, the FDR depends on the distribution of the test statistics under the alternative hypotheses H'_j, for $j = m_0 + 1, \ldots, m$, through the random variable S (here, the FDR is a function of β, the rejection probability under the alternative hypotheses). In general, the FDR is thus more difficult to work with than the other three error rates discussed so far. Figure 1 displays plots of the FWER, PCER and FDR versus the number of hypotheses m, for different proportions $m_0/m = 1, 0.9, 0.5, 0.1$ 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 in the figure because it is on a different scale, that is, it is not restricted to belong to the interval $[0, 1]$). 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. We plotted the error rates for values of m between 1 and 100 only to provide more detail in regions where there are sharp changes in these error rates. For larger m's, in the thousands as in DNA microarray experiments, the error rates tend to reach a plateau. Figure 2 displays plots of the FWER, PCER and FDR versus 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 again becomes closer to the PCER. The error rates display similar behavior for larger values of the number of hypotheses m, with a sharper increase of the FWER as α increases.

2.3 *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 (1) $\Pr(T_1 \in S_\alpha \mid \mathrm{H}_1) = \alpha$ for all $\alpha \in [0, 1]$ which are achievable under the distribution of T_1 and (2) $S_{\alpha'} = \bigcap_{\alpha \geq \alpha'} S_\alpha$ for all α and α' for which these regions are defined in (1). Rather than simply reporting rejection or nonrejection of the hypothesis H_1, a *p-value* connected with the test can be defined as $p_1 = \inf\{\alpha : t_1 \in S_\alpha\}$ (adapted from Lehmann, 1986, page 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, given t_1. 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 be restated as the probability of observing a test statistic as extreme or more extreme in

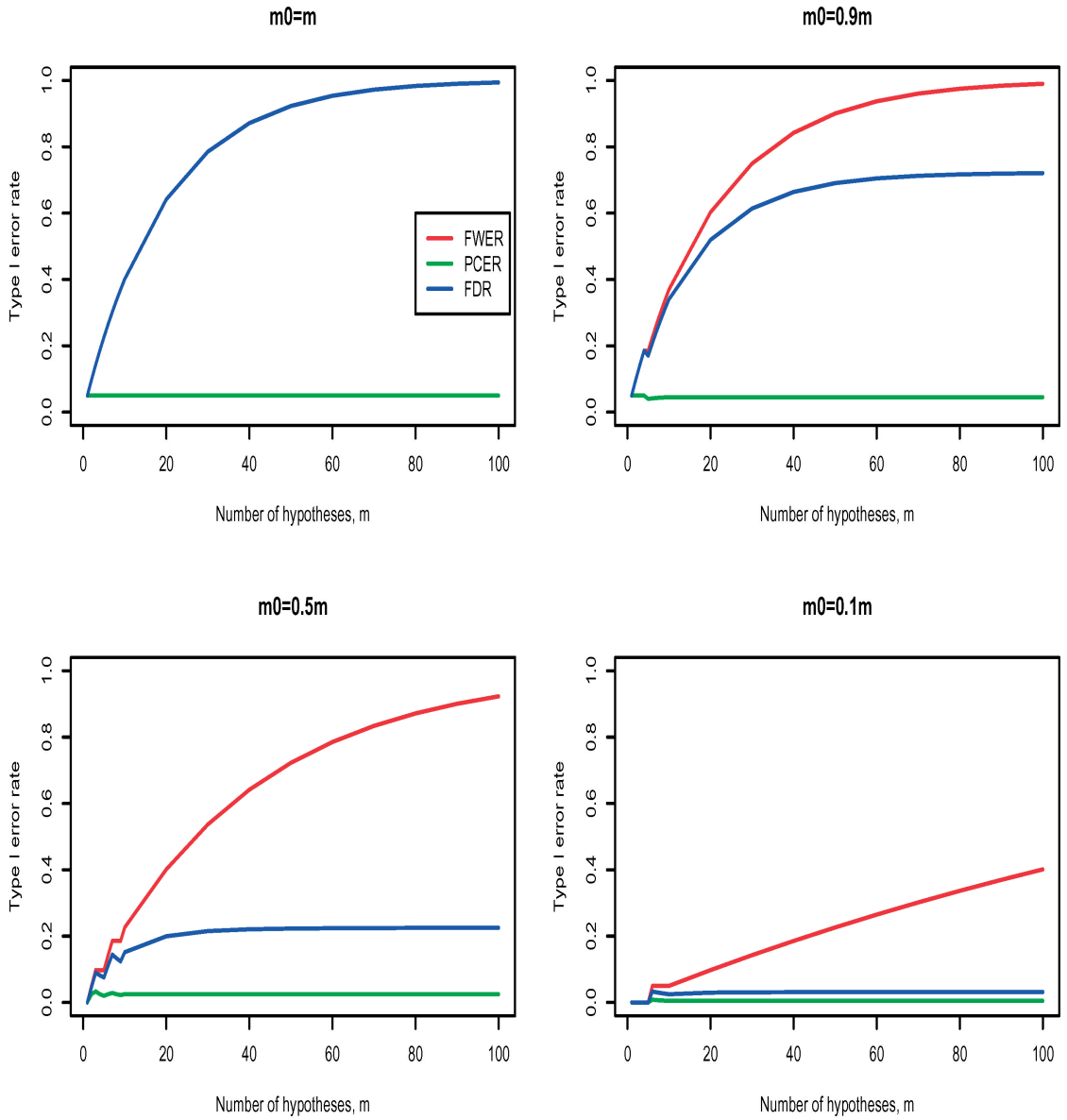

FIG. 1. *Type I error rates, simple example. Plot of Type I error rates versus number of hypotheses m for different proportions of true null hypotheses,* $m_0/m = 1, 0.9, 0.5, 0.1$. *The model and multiple testing procedures are described in Section* 2.2. *The individual test size is* $\alpha = 0.05$ *and the parameter d was set to* 1. *The nonsmooth behavior for small m is due to the fact that it is not always possible to have exactly* 90, 50, *or* 10% *of true null hypotheses and rounding to the nearest integer is necessary.* FWER: *red curve;* FDR: *blue curve;* PCER: *green curve.*

the direction of rejection as the observed one, that is, $p_1 = \Pr(|T_1| \geq |t_1| \mid H_1)$. Extending the concept of p-value to the multiple testing situation leads to the very useful definition of adjusted p-value.

Adjusted p-values. Let t_j and $p_j = \Pr(|T_j| \geq |t_j| \mid H_j)$ denote, respectively, the test statistic and *unadjusted* or *raw* 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: for example, 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 described in terms of adjusted p-values. Given any test procedure, the *adjusted p-value* that corresponds to the test of a single hypothesis H_j can be defined as the nominal level of the entire test procedure at which H_j

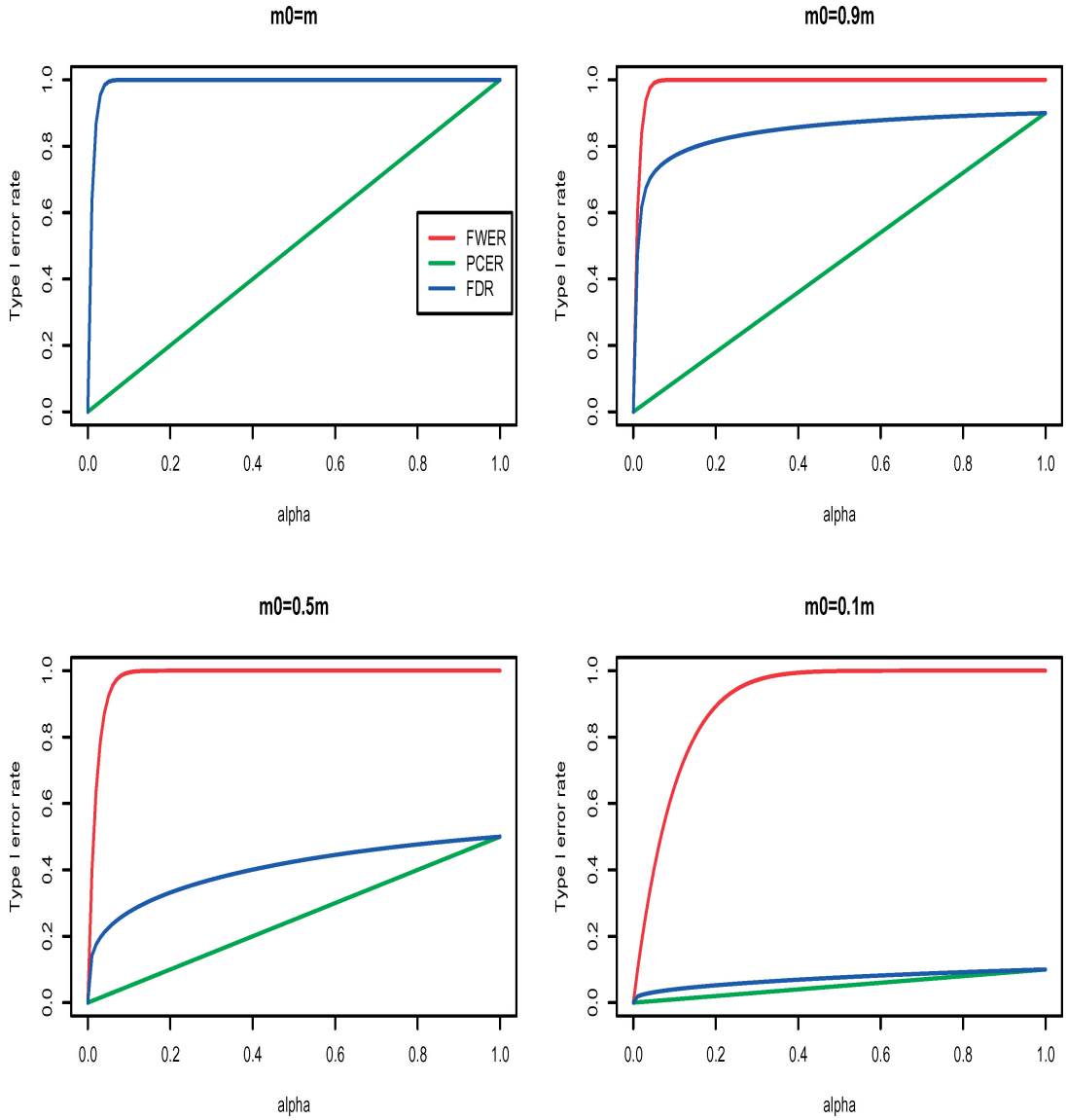

FIG. 2. *Type I error rates, simple example. Plot of Type I error rates versus individual test size* α*, for different proportions of true null hypotheses,* $m_0/m = 1, 0.9, 0.5, 0.1$*. The model and multiple testing procedures are described in Section* 2.2*. The number of hypotheses is* $m = 100$ *and the parameter* d *was set to* 1. FWER: *red curve;* FDR: *blue curve;* PCER: *green curve.*

would just be rejected, given the values of all test statistics involved (Hommel and Bernhard, 1999; Shaffer, 1995; Westfall and Young, 1993; Wright, 1992; Yekutieli and Benjamini, 1999). If interest is in controlling the FWER, the adjusted p-value for hypothesis H_j, given a specified multiple testing procedure, is $\tilde{p}_j = \inf\{\alpha \in [0,1] : H_j \text{ is rejected at nominal FWER} = \alpha\}$, where the *nominal* FWER is the α-level at which the specified procedure is performed. The corresponding random variables for unadjusted and adjusted p-values are denoted by P_j and $\tilde{P}_j$, respectively. Hypothesis H_j is then rejected, that is, gene j is declared differentially expressed at nominal FWER α if $\tilde{p}_j \leq \alpha$. Note that for many procedures, such as the Bonferroni procedure described in Section 2.4.1, the *nominal* level is usually larger than the *actual* level, thus resulting in a conservative test. Adjusted p-values for procedures controlling other types of error rates are defined similarly, that is, for FDR controlling procedures, $\tilde{p}_j = \inf\{\alpha \in [0,1] : H_j \text{ is rejected at nominal FDR} = \alpha\}$ (Yekutieli and 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 estimated by resampling methods (Westfall and 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 that correspond to the *most significant* test statistics (i.e., smallest unadjusted p-values or largest absolute test statistics) are considered successively, with further tests dependent on the outcomes of earlier ones. As soon as one fails to reject a null hypothesis, no further hypotheses are rejected. In contrast, for *step-up* procedures, the hypotheses that correspond to the *least significant* test statistics are considered successively, again with further tests dependent 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 unadjusted p-value less than or equal to α/m. The corresponding *single-step Bonferroni adjusted p-values* are thus given by

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

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

$$\begin{aligned}\mathrm{FWER} &= \Pr(V \geq 1) \\ &= \Pr\left(\bigcup_{j=1}^{m_0} \{\tilde{P}_j \leq \alpha\}\right) \leq \sum_{j=1}^{m_0} \Pr(\tilde{P}_j \leq \alpha) \\ &\leq \sum_{j=1}^{m_0} \Pr\left(P_j \leq \frac{\alpha}{m}\right) \leq \frac{m_0\alpha}{m},\end{aligned}$$

where the last inequality follows from $\Pr(P_j \leq x \mid H_j) \leq x$, for any $x \in [0, 1]$.

Closely related to the Bonferroni procedure is the Šidák procedure. It is exact for protecting the FWER under the complete null, 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 unadjusted p-values, are dependent. This is the case in DNA microarray experiments, where groups of genes tend to have highly correlated expression measures due, for example, to co-regulation. Westfall and Young (1993) proposed adjusted p-values for less conservative multiple testing procedures which take into account the dependence structure among test statistics. The *single-step min P adjusted p-values* are defined by

$$\tilde{p}_j = \Pr\left(\min_{1 \leq l \leq m} P_l \leq p_j \mid \mathrm{H}_0^C\right), \tag{3}$$

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

$$\tilde{p}_j = \Pr\left(\max_{1 \leq l \leq m} |T_l| \geq |t_j| \mid \mathrm{H}_0^C\right). \tag{4}$$

The following points should be noted regarding the four procedures introduced above.

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 min P 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: $\Pr(|T_1| \leq c_1, \ldots, |T_m| \leq c_m) \geq \prod_{j=1}^{m} \Pr(|T_j| \leq c_j)$. This inequality, also known as the *positive orthant dependence property*, was initially derived by Dunn (1958) for $(T_1, \ldots, T_m)$ that have a multivariate normal distribution with mean zero and certain types of covariance matrices. Šidák (1967) extended the result to arbitrary covariance matrices and Jogdeo (1977) showed that the inequality holds for a larger class of distributions, including some multivariate t- and F-distributions.

When the Šidák inequality holds, the $\min P$ 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. In other words, procedures based on the $\min P$ 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 $\min P$ adjustments are equivalent as discussed in item 1, above.
4. Procedures based on the $\max T$ and $\min P$ adjusted p-values provide *weak* control of the FWER. *Strong* control of the FWER holds under the assumption of subset pivotality (Westfall and Young, 1993, page 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 random vector $\{P_j : j \in \Lambda_0\}$ is identical for distributions satisfying the composite null hypotheses $\bigcap_{j\in\Lambda_0} \mathrm{H}_j$ and $\mathrm{H}_0^C = \bigcap_{j=1}^m \mathrm{H}_j$, for all subsets Λ_0 of $\{1, \ldots, m\}$. Here, composite hypotheses of the form $\bigcap_{j\in\Lambda_0} \mathrm{H}_j$ refer to the *joint* distribution of test statistics T_j or p-values P_j for testing hypotheses H_j, $j \in \Lambda_0$. Without subset pivotality, multiplicity adjustment is more complex, as one would need to consider the distribution of the test statistics under partial null hypotheses $\bigcap_{j\in\Lambda_0} \mathrm{H}_j$, rather than the complete null hypothesis H_0^C. In the microarray context considered in this article, each null hypothesis refers to a single gene j and each test statistic T_j is a function of the response/covariate Y and expression measure X_j only. The composite hypothesis $\bigcap_{j\in\Lambda_0} \mathrm{H}_j$ refers to the *joint* distribution of variables Y and $\{X_j : j \in \Lambda_0\}$ and specifies that the random subvector of expression measures $\{X_j : j \in \Lambda_0\}$ is independent of the response/covariate Y, i.e., that the *joint* distribution of $\{X_j : j \in \Lambda_0\}$ is identical for all levels of Y. The pivotality property holds given the assumption that test statistics for genes in the null subset Λ_0 have the same *joint* distribution regardless of the truth or falsity of the hypotheses in the complement of Λ_0. For a discussion of subset pivotality and examples of testing problems in which the condition holds and does not hold, see Westfall and Young (1993).
5. The $\max T$ adjusted p-values are easier to compute than the $\min P$ p-values and are equal to the $\min P$ 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 $\max T$ adjusted p-values and this can lead to unbalanced adjustments (Beran, 1988; Westfall and Young, 1993, page 50). When adjusted p-values are estimated by permutation (Section 2.6) and a large number of hypotheses are tested, procedures based on the $\min P$ p-values tend to be more sensitive to the number of permutations and more conservative than those based on the $\max T$ p-values. Also, the $\min P$ procedure requires more computations than the $\max T$ procedure, because the unadjusted p-values must be estimated before considering the distribution of their successive minima (Ge, Dudoit and Speed, 2003).

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 $p_{r_1} \le p_{r_2} \le \cdots \le p_{r_m}$ denote the *observed ordered unadjusted p-values* and let $\mathrm{H}_{r_1}, \mathrm{H}_{r_2}, \ldots, \mathrm{H}_{r_m}$ denote the corresponding null hypotheses. For strong control of the FWER at level α, the Holm (1979) procedure proceeds as follows. Define $j^* = \min\{j : p_{r_j} > \alpha/(m - j + 1)\}$ 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

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

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

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

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

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

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

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

where $|t_{r_1}| \geq |t_{r_2}| \geq \cdots \geq |t_{r_m}|$ denote the *observed ordered test statistics*. Note that applying Boole's inequality to the quantities in (7) yields Holm's p-values. A procedure based on the step-down min P adjusted p-values is thus less conservative than Holm's procedure. For a proof of the strong control of the FWER for the max T and min P procedures the reader is referred to Westfall and 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 $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ satisfy

$$\Pr\left(P_{(j)} > \frac{\alpha j}{m},\ \forall j = 1,\ldots,m \,|\, \mathrm{H}_0^C \right) \geq 1 - \alpha$$

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.

Hochberg (1988) used the Simes inequality to derive the following FWER controlling procedure. For control of the FWER at level α, let $j^* = \max\{j : p_{r_j} \leq \alpha/(m - j + 1)\}$ 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

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

The Hochberg (1988) procedure can be viewed as the step-up analog of Holm's step-down procedure, since the ordered unadjusted p-values are compared to the same critical values in both procedures, namely, $\alpha/(m - j + 1)$. Related procedures include those of Hommel (1988) and Rom (1990). Step-up procedures often have 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 DNA 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 and Young (1993) step-down max T procedure.

2.5 Control of the False Discovery Rate

A different approach to multiple testing was proposed in 1995 by Benjamini and Hochberg. These authors argued 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. Specifically, the FDR is defined as FDR $= E(Q)$, where $Q = V/R$ if $R > 0$ and 0 if $R = 0$, that is, 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 that control 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 and Hochberg (1995) derived the following step-up procedure for (strong) control of the FDR for independent test statistics. Let $p_{r_1} \leq p_{r_2} \leq \cdots \leq p_{r_m}$ denote the observed ordered unadjusted p-values. For control of the FDR at level α define $j^* = \max\{j : p_{r_j} \leq (j/m)\alpha\}$ and reject hypotheses H_{r_j} for $j = 1,\ldots,j^*$. If no such j^* exists, reject no hypothesis. Corresponding adjusted p-values are

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

Benjamini and Yekutieli (2001) proved that this procedure controls the FDR under certain dependence structures (for example, positive regression dependence). They also proposed a simple conservative modification

of the procedure which controls the false discovery rate for arbitrary dependence structures. Adjusted p-values for the modified step-up procedure are

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

The above two step-up procedures differ only in their penalty for multiplicity, that is, in the multiplier applied to the unadjusted p-values. For the standard Benjamini and Hochberg (1995) procedure, the penalty is m/k [Equation (10)], while for the conservative Benjamini and Yekutieli (2001) procedure it is $(m\sum_{j=1}^{m}1/j)/k$ [Equation (11)]. For a large number m of hypotheses, the penalties differ by a factor of about $\log m$. Note that the Benjamini and Hochberg procedure can be conservative even in the independence case, as it was shown that for this step-up procedure $E(Q) \leq (m_0/m)\alpha \leq \alpha$. Until recently, most FDR controlling procedures were either designed for independent test statistics or did not make use of the dependence structure among the test statistics. In the spirit of the Westfall and Young (1993) resampling procedures for FWER control, Yekutieli and Benjamini (1999) proposed new FDR controlling procedures that use resampling-based adjusted p-values to incorporate certain types of dependence 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). Other recent work on FDR controlling procedures can be found in Genovese and Wasserman (2001), Storey (2002), and Storey and Tibshirani (2001).

In the microarray setting, where thousands of tests are performed simultaneously and a fairly large number of genes are expected to be differentially expressed, FDR controlling procedures present a promising alternative to 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 even marginal) distribution of the test statistics is unknown. Resampling methods (e.g., bootstrap, permutation) can be used to estimate unadjusted and adjusted p-values while avoiding parametric assumptions about the joint distribution of the test statistics. Here, we consider null hypotheses H_j of no association between variable X_j and a response or covariate Y, $j = 1, \ldots, m$. In the microarray setting and for this type of null hypothesis, the joint distribution of the test statistics $(T_1, \ldots, T_m)$ under the complete null hypothesis can be estimated by permuting the columns of the gene expression data matrix X (see Box 1). Permuting entire columns of this matrix creates a situation in which the response or covariate Y is independent of the gene expression measures, while attempting to preserve the correlation structure and distributional characteristics of the gene expression measures. Depending on the sample size n, it may be infeasible to consider all possible permutations; for large n, a random subset of B permutations (including the observed) may be considered. The manner in which the responses/covariates are permuted should reflect 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 Section 9.3 in Scheffé (1959) and Section 3.1.2 in the present article].

Note that while permutation is appropriate for the types of null hypotheses considered in this article, permutation procedures are not advisable for certain other types of hypotheses. For instance, consider the simple case of a binary variable $Y \in \{1, 2\}$ and suppose that the null hypothesis H_j is that the conditional distributions of X_j given $Y = 1$ and of X_j given $Y = 2$ have equal means, but possibly different variances. A permutation null distribution enforces *equal distributions* in the two groups, which is clearly stronger

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 (i.e., gene).

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

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

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

than simply equal means. As a result, a null hypothesis H_j may be rejected for reasons other than a difference in means (e.g., difference in a nuisance parameter). Bootstrap resampling is more appropriate for this type of hypotheses, as it preserves the covariance structure present in the original data. The reader is referred to Pollard and van der Laan (2003) for a discussion of resampling-based methods in multiple testing.

Permutation adjusted p-values for the Bonferroni, Šidák, Holm and Hochberg procedures can be obtained by replacing p_j by p_j^* (see Box 1) 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 $\max T$ adjusted p-values (see Box 2) of Westfall and Young (1993), the complete 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 is omitted here; in that case, one needs only the distribution of the maximum $\max_{1\le l\le m}|T_l|$.)

The reader is referred to Ge, Dudoit and Speed (2003) for a fast permutation algorithm for estimating $\min P$ adjusted p-values.

Box 2. Permutation algorithm for step-down $\max T$ adjusted p-values based on Algorithms 2.8 and 4.1 in Westfall and 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 (i.e., gene).
3. Next, compute successive maxima of the test statistics

$$u_{m,b} = |t_{r_m,b}|,$$

$$u_{j,b} = \max(u_{j+1,b}, |t_{r_j,b}|) \quad \text{for } j = m-1, \ldots, 1,$$

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

The permutation adjusted p-values are

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

with the monotonicity constraints enforced by setting

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

2.7 Recent Proposals for DNA Microarray Experiments

Efron et al. (2000), Golub et al. (1999), and Tusher, Tibshirani and Chu (2001) have recently proposed resampling algorithms for multiple testing in DNA microarray experiments. However, these oft-cited 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 types of leukemias: acute lymphoblastic leukemia (ALL, class 1) and acute myeloid leukemia (AML, class 2). (The study is described in greater detail in Section 3.1.3.) In their so-called *neighborhood analysis*, the authors computed a test statistic t_j for each gene [$P(g, c)$ in their notation],

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

where $\bar{x}_{kj}$ and s_{kj} denote, respectively, the average and standard deviation of the expression measures of gene j in the class $k = 1, 2$ samples. The Golub et al. 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., 1999). It is not pivotal, even in the Gaussian case or asymptotically, 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: here, the statistic is said to be pivotal if its null distribution does not depend on parameters of the distribution which generated the data.) Statistics such as the Golub et al. statistic above have been used in meta-analysis to measure effect sizes (National Reading Panel, 1999).

Golub et al. used the term *neighborhood* to refer to sets of genes with test statistics T_j greater in absolute value than a given critical value $c > 0$, that is, sets of rejected hypotheses $\{j : T_j \ge c\}$ or $\{j : T_j \le -c\}$ [these sets are denoted by $N_1(c, r)$ and $N_2(c, r)$, respectively, in note 16 of Golub et al., 1999]. The ALL/AML labels were permuted $B = 400$ times to estimate the complete null distribution of the numbers $R(c) = V(c) = \sum_{j=1}^{m} I(T_j \ge c)$ of false positives for different critical values c (similarly for the other tail, with $T_j \le -c$). Figure 2 in Golub et al. (1999) contains plots

of the observed $R(c) = r(c)$ and permutation quantiles of $R(c)$ against critical values c for one-sided tests. [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 the Golub et al. method note that we use T_j to denote $P(g, c)$, c to denote r and $r(c)$ to denote a realization of $R(c)$, that is, $|N_1(c, r)| + |N_2(c, r)|$.] 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, $G(c) = \Pr(R(c) \geq r(c) \mid \mathrm{H}_0^C) = \alpha$.

Golub et al. provided 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, $G(c) = \Pr(R(c) \geq r(c) \mid \mathrm{H}_0^C)$. $G(c)$ 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. Dudoit, Shaffer and Boldrick (2002) show that conditional on the observed ordered absolute test statistics, $|t|_{(1)} \geq \cdots \geq |t|_{(m)}$, the function $G(c)$ is left-continuous with discontinuities at $|t|_{(j)}$, $j = 1, \ldots, m$. 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 the step-down and step-up procedures described in Dudoit, Shaffer and Boldrick (2002). It turns out that neither version provides strong control of any Type I error rate. The step-down version does, however, control the FWER weakly. Finally, 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 $G(c) = \Pr(R(c) \geq r(c) \mid \mathrm{H}_0^C)$ versus the critical values c, because this does not require a prespecified level α. A more detailed discussion of the statistical properties of neighborhood analysis and related figures can be found in Dudoit, Shaffer and Boldrick (2002).

2.7.2 *Significance Analysis of Microarrays.* We consider the significance analysis of microarrays (SAM) multiple testing procedure described in Tusher, Tibshirani and Chu (2001) and Chu et al. (2000). An earlier version of SAM (Efron et al., 2000) is discussed in detail in the technical report by Dudoit, Shaffer and Boldrick (2002).* Note that the SAM articles 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 multiple testing problem and assume that a suitable test statistic is computed for each gene.

SAM procedure from Tusher, Tibshirani and Chu (2001).

1. Compute a test statistic t_j for each gene j and define order statistics $t_{(j)}$ such that $t_{(1)} \geq t_{(2)} \geq \cdots \geq t_{(m)}$. [The notation for the ordered test statistics is different here than in Tusher, Tibshirani and Chu (2001) to be consistent with previous notation whereby we set $t_{(1)} \geq t_{(2)} \geq \cdots \geq t_{(m)}$ and $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.]
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 $t_{(1),b} \geq t_{(2),b} \geq \cdots \geq t_{(m),b}$. Note that $t_{(j),b}$ may correspond to a different gene than the observed $t_{(j)}$.
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 (Q–Q) plot (so-called SAM plot) of the observed $t_{(j)}$ versus the expected $\bar{t}_{(j)}$.
5. 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 $j_2 = \min\{j > j_0 : t_{(j)} - \bar{t}_{(j)} \leq -\Delta\}$. [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.] 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, $\mathrm{cut}_{\mathrm{up}}(\Delta) = \min\{t_{(j)} : j \leq j_1\} = t_{(j_1)}$, and the lower cut point, $\mathrm{cut}_{\mathrm{low}}(\Delta) = \max\{t_{(j)} : j \geq j_2\} = t_{(j_2)}$. If no such j_1 (j_2) exists, set $\mathrm{cut}_{\mathrm{up}}(\Delta) = \infty$ ($\mathrm{cut}_{\mathrm{low}}(\Delta) = -\infty$).
6. For a given threshold Δ, the expected number of false positives, PFER, is estimated by computing for each of the B permutations the number of genes with $t_{j,b}$ above $\mathrm{cut}_{\mathrm{up}}(\Delta)$ or below $\mathrm{cut}_{\mathrm{low}}(\Delta)$, and averaging this number over permutations.

*See "Note Added in Proof" on p. 101.

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

Thus, the SAM procedure in Tusher, Tibshirani and Chu (2001) uses ordered test statistics from the original data only for the purpose of obtaining global cutoffs for the test statistics. In the permutation, the cutoffs are kept fixed and the PFER is estimated by counting the number of genes with test statistics above/below these global cutoffs. Note that the cutoffs are actually random variables, because they depend on the observed test statistics.

The reader is referred to Dudoit, Shaffer and Boldrick (2002) for a more detailed discussion and comparison of the statistical properties of the SAM procedures in Efron et al. (2000) and Tusher, Tibshirani and Chu (2001), including a derivation of the corresponding adjusted p-values and an extension which accounts for differences in variances among the order statistics. There, it is shown that both SAM procedures aim to control the PFER (or PCER), but the Efron et al. (2000) procedure controls this error rate only in the weak sense. The only difference between the Tusher, Tibshirani and Chu (2001) 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 (a discussion of asymmetric critical values is found in Braver, 1975). 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. (2000) amounts to rejecting $H_{(j)}$ 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, Tibshirani and Chu (2001) rejects H_j whenever $t_j \geq \text{cut}_{\text{up}}(\Delta)$ or $t_j \leq \text{cut}_{\text{low}}(\Delta)$, where $\text{cut}_{\text{low}}(\Delta)$ and $\text{cut}_{\text{up}}(\Delta)$ are chosen from the permutation quantile–quantile plot and such that the PFER is controlled strongly at a given level.

SAM control of the FDR. The term "false discovery rate" is misleading, as the definition in SAM is different than the standard definition of Benjamini and Hochberg (1995): the SAM FDR is estimating $E(V \mid H_0^C)/R$ and not $E(V/R)$ as in Benjamini and Hochberg. Furthermore, the FDR in SAM can be greater than 1 (cf. Table 3 in Chu et al., 2000, page 16). The issue of strong versus weak control is only mentioned briefly in Tusher, Tibshirani and Chu (2001) and the authors claim that "SAM provides a reasonably accurate estimate for the true FDR."

2.8 Reporting the Results of Multiple Testing Procedures

We have described a number of multiple testing procedures for controlling different Type I error rates, including the FWER and the FDR. Table 2 summarizes these methods in terms of their main properties: definition of the Type I error rate, type of control of this error rate (strong versus weak), stepwise nature of the procedure, distributional assumptions.

For each procedure, adjusted p-values were derived as convenient and flexible summaries of the strength of the evidence against each null hypothesis. The following types of plots of adjusted p-values are particularly useful in summarizing the results of different multiple testing procedures applied to a large number of genes. The plots allow biologists to examine various false positive rates (FWER, FDR, or PCER) associated with different gene lists. They do not require researchers to preselect a particular definition of Type I error rate or α-level, but rather provide them with tools to decide on an appropriate combination of number of genes and tolerable false positive rate for a particular experiment and available resources.

Plot of ordered adjusted p-values ($\tilde{p}_{(j)}$ versus j). For a given number of genes r, say, this representation provides the nominal Type I error rate for a given procedure when the r genes with the smallest adjusted p-values for that procedure are declared to be differentially expressed [see panels (a) and (b) in Figures 3–5]. Therefore, rather than choosing a specific type of error control and α-level, biologists might first select a number r of genes which they feel comfortable following up. The nominal false positive rates (or adjusted p-values, $\tilde{p}_{(r)}$) corresponding to this number under various types of error control and procedures can then be read from the plot. For instance, for $r = 10$ genes, the nominal FWER from Holm's step-down procedure might be 0.1 and the nominal FDR from the Benjamini and Hochberg (1995) step-up procedure might be 0.07.

Plot of number of genes declared to be differentially expressed versus nominal Type I error rate (r versus α). This type of plot is the "transpose" of the previous plot and can be used as follows. For a given nominal level α, find the number r of genes that would be declared to be differentially expressed under one procedure, and read the level required to achieve that number under other methods. Alternatively, find the number of genes that would be identified using a procedure

TABLE 2
Properties of multiple testing procedures

Procedure	Type I error rate	Strong or weak control	Stepwise structure	Dependence structure
Bonferroni	FWER	Strong	Single	General/ignore
Šidák	FWER	Strong	Single	Positive orthant dependence
min P	FWER	Strong	Single	Subset pivotality
max T	FWER	Strong	Single	Subset pivotality
Holm (1979)	FWER	Strong	Down	General/ignore
Step-down Šidák	FWER	Strong	Down	Positive orthant dependence
Step-down min P	FWER	Strong	Down	Subset pivotality
Step-down max T	FWER	Strong	Down	Subset pivotality
Hochberg (1988)	FWER	Strong	Up	Some dependence (Simes)
Troendle (1996)	FWER	Strong	Up	Some dependence
Benjamini and Hochberg (1995)	FDR	Strong	Up	Positive regression dependence
Benjamini and Yekutieli (2001)	FDR	Strong	Up	General/ignore
Yekutieli and Benjamini (1999)	FDR	Strong	Up	Some dependence
Unadjusted p-values	PCER	Strong	Single	General/ignore
SAM, Tusher, Tibshirani and Chu (2001)	PFER (PCER)	Strong	Single	General/hybrid
SAM, Efron et al. (2000)	PFER (PCER)	Weak	Single	General
Golub et al. (1999), step-down	$\Pr(R \geq r \mid \mathrm{H}_0^C)$ (FWER)	Weak	Down	General
Golub et al. (1999), step-up	$\Pr(R \geq r \mid \mathrm{H}_0^C)$	Weak	Up	General

Notes. By "General/ignore," we mean that a procedure controls the claimed Type I error rate for general dependence structures, but does not explicitly take into account the joint distribution of the test statistics. For the Tusher, Tibshirani and Chu (2001) SAM version, the term "General/hybrid" refers to the fact that only the marginal distribution of the test statistics is considered when computing the PFER. The test statistics are considered jointly only to determine the cutoffs $\mathrm{cut}_{\mathrm{up}}(\Delta)$ and $\mathrm{cut}_{\mathrm{low}}(\Delta)$ from the quantile–quantile plot.

controlling the FWER at a fixed nominal level α, and then identify how many others would be identified using procedures controlling the FDR and PCER at that level.

The multiple testing procedures considered in this article can be divided into the following two broad categories: those for which adjusted p-values are monotone in the test statistics, t_j, and those for which adjusted p-values are monotone in the unadjusted p-values, p_j. In general, the ordering of genes based on test statistics t_j will differ from that based on unadjusted p-values p_j, as the test statistics of different genes may have different distributions. Within each of these two classes, procedures will, however, produce the same orderings of genes, whether they are designed to control the FWER, FDR or PCER. They will differ only in the cutoffs for significance. That is, for a given nominal level α, an FWER controlling procedure such as Bonferroni's might identify only the first 20 genes with the smallest unadjusted p-values, while an FDR controlling procedure such as Benjamini and Hochberg's (1995) might retain an additional 15 genes with the next 15 smallest unadjusted p-values.

3. DATA

3.1 Microarray Data

3.1.1 *Apolipoprotein AI experiment of Callow et al.* The apolipoprotein AI (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 is a gene known to play a pivotal role in high density lipoprotein (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 apo AI knock out mice compared to inbred control mice. The treatment group consisted of eight inbred C57Bl/6 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

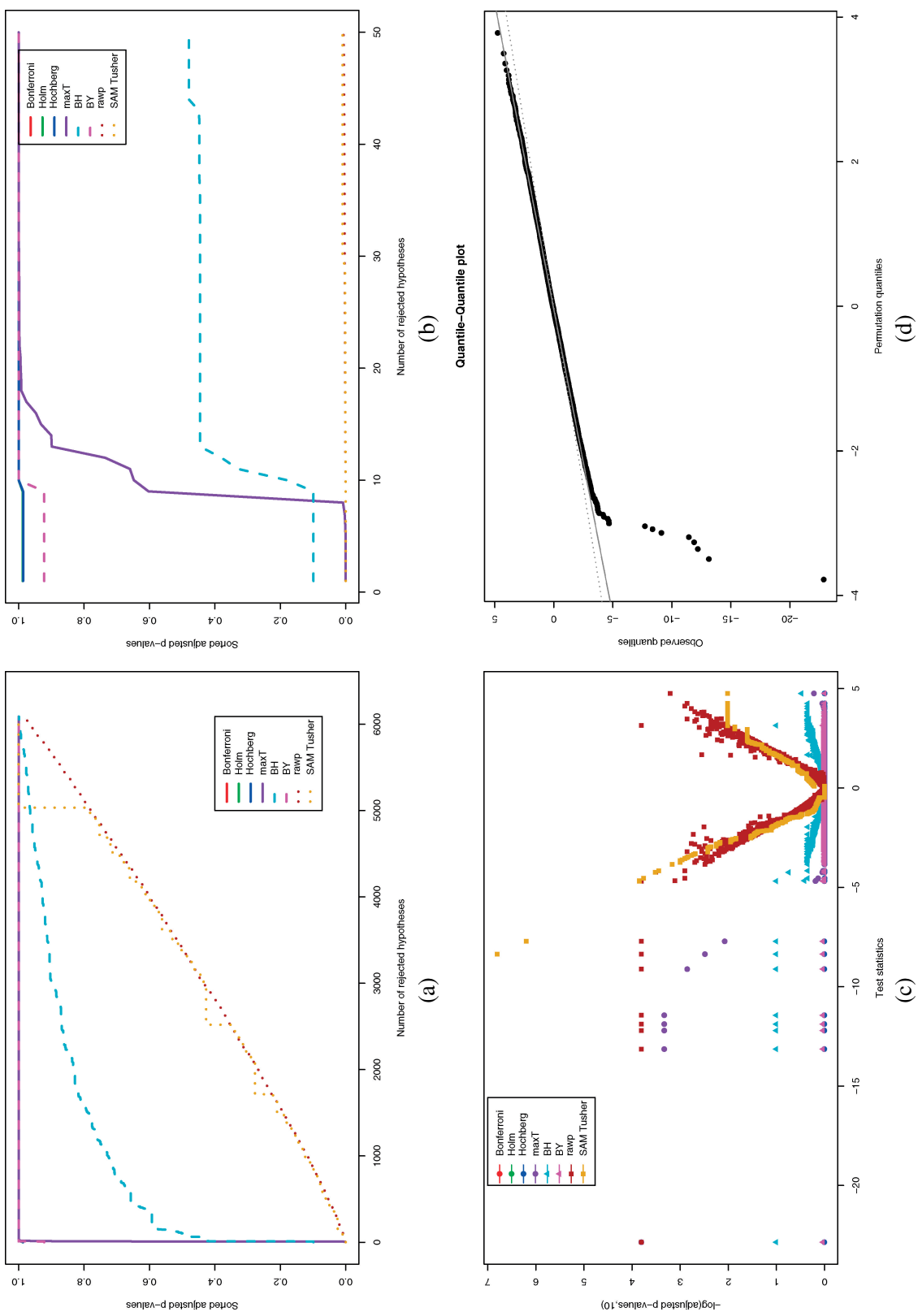

FIG. 3. *Apo AI experiment.* (a) *and* (b) *Plot of sorted permutation adjusted p-values,* $\tilde{p}^*_{(j)}$ *versus* j. *Panel* (b) *is an enlargement of panel* (a) *for the* 50 *genes with the largest absolute t-statistics* $|t_i|$. *Adjusted p-values for procedures controlling the* FWER, FDR *and* PCER *are plotted using solid, dashed and dotted lines, respectively.* (c) *Plot of adjusted p-values,* $-\log_{10}\tilde{p}^*_j$ *versus t-statistics* t_j. (d) *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* $B_{\text{perm}} = \binom{16}{8} = 12{,}870$ *permutations of the treatment/control labels, except for the SAM procedure for which* $B_{\text{sam}} = 1000$ *random permutations were used. Note that the results for the Bonferroni, Holm, and Hochberg procedures are virtually identical; similarly for the unadjusted p-value (rawp) and SAM Tusher procedures.*

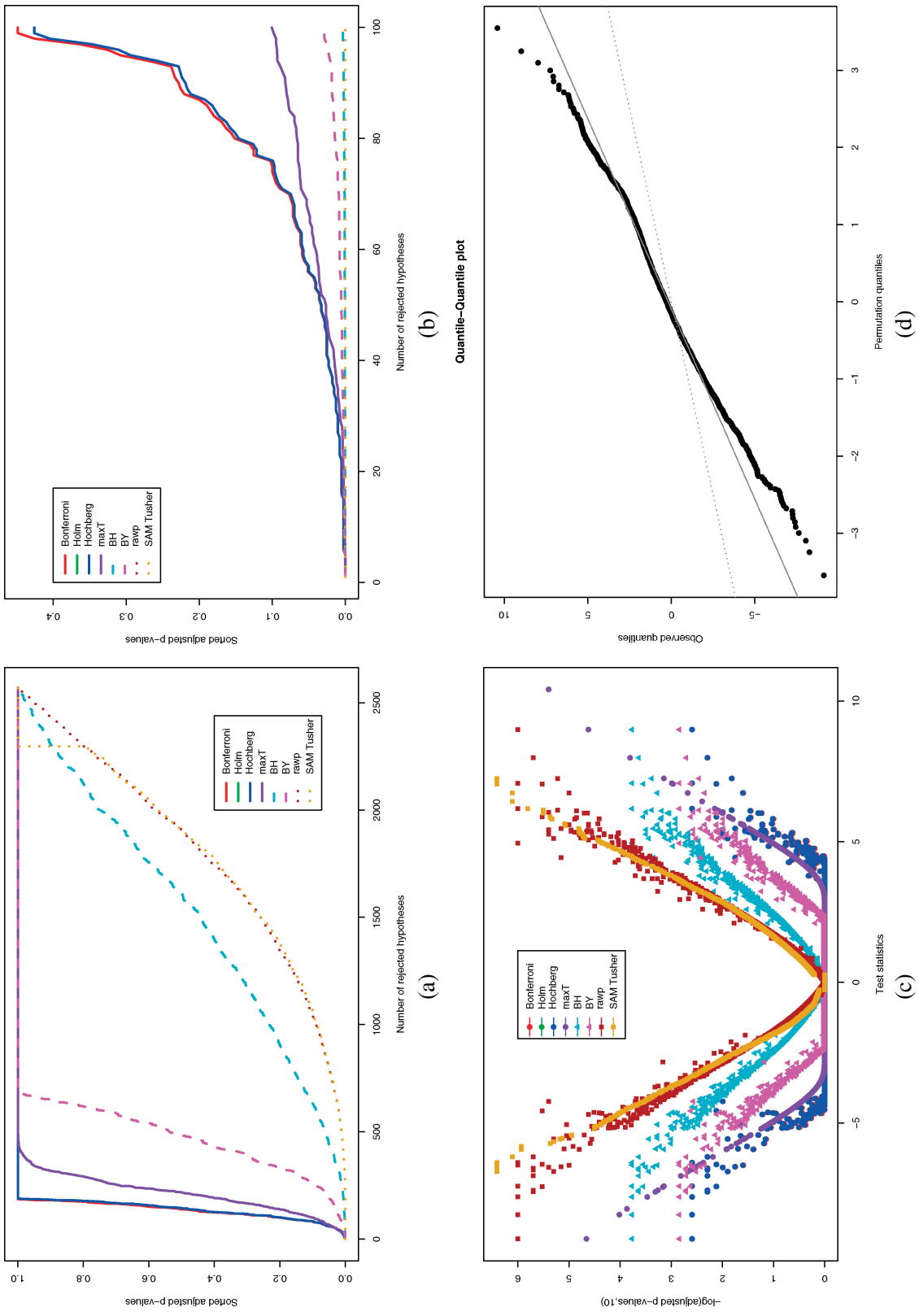

FIG. 4. *Bacteria experiment.* (a) *and* (b) *Plot of sorted permutation adjusted p-values,* $\tilde{p}^*_{(j)}$ *versus* j. *Panel* (b) *is an enlargement of panel* (a) *for the* 100 *genes with the largest absolute t-statistics* $|t_j|$. *Adjusted p-values for procedures controlling the* FWER, FDR *and* PCER *are plotted using solid, dashed and dotted lines, respectively.* (c) *Plot of adjusted p-values,* $-\log_{10}\tilde{p}^*_j$ *versus t-statistics* t_j. (d) *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* $B_{\text{perm}} = 2^{22}$ *permutations of the Gram-positive/Gram-negative labels within the* 22 *dose* $\times$ *time blocks, except for the SAM procedure for which* $B_{\text{sam}} = 1000$ *random permutations were used. Note that the results for the Bonferroni, Holm, and Hochberg procedures are virtually identical, similarly for the unadjusted p-value (rawp) and SAM Tusher procedures.*

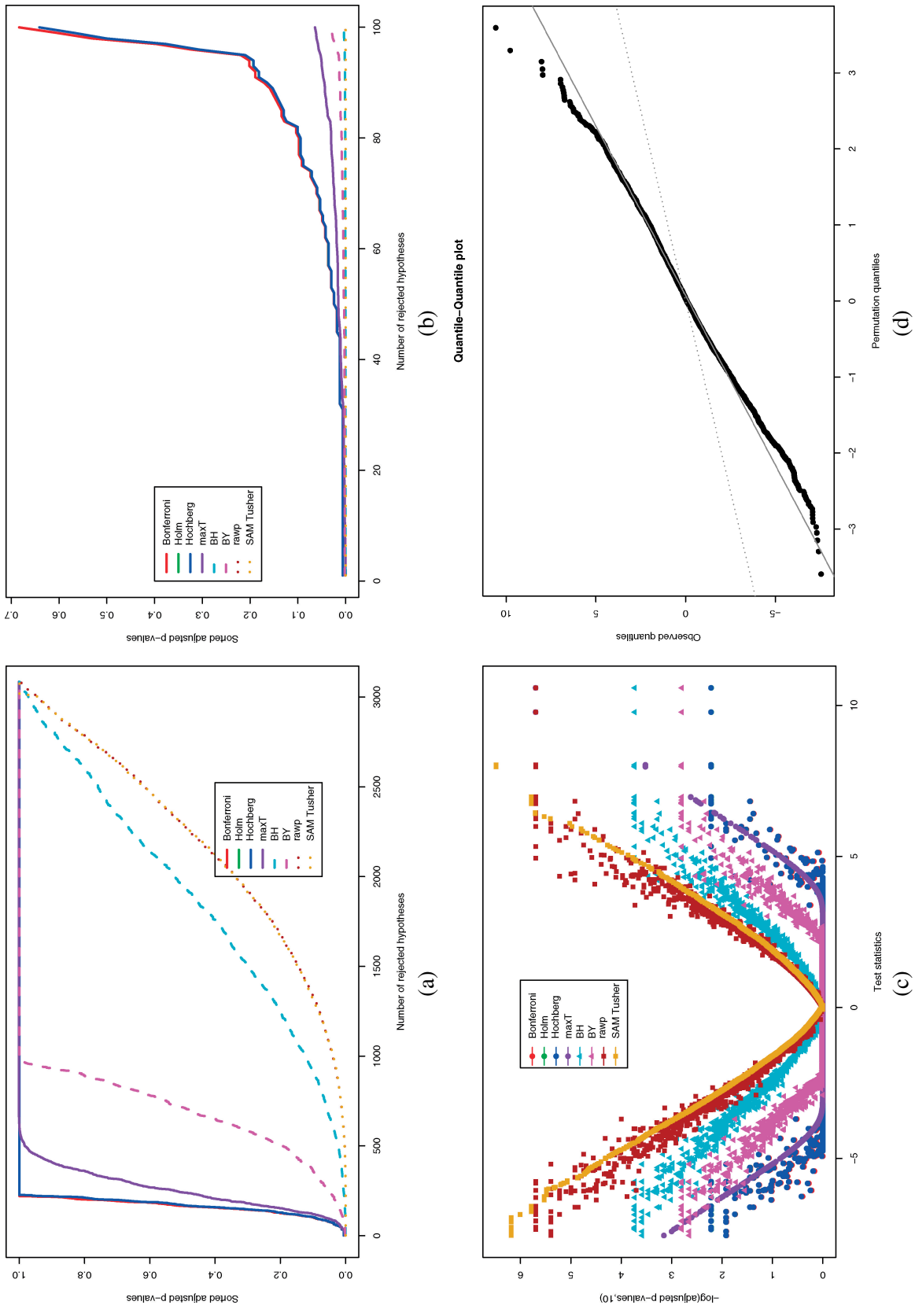

FIG. 5. *Leukemia study.* (a) *and* (b) *Plot of sorted permutation adjusted p-values,* $\tilde{p}^*_{(j)}$ *versus* j. *Panel* (b) *is an enlargement of panel* (a) *for the* 100 *genes with the largest absolute t-statistics* $|t_j|$. *Adjusted p-values for procedures controlling the* FWER, FDR *and* PCER *are plotted using solid, dashed and dotted lines, respectively.* (c) *Plot of adjusted p-values,* $-\log_{10}\tilde{p}^*_j$ *versus t-statistics* t_j. (d) *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* $B_{\text{perm}} = 500{,}000$ *random permutations of the ALL/AML labels, except for the SAM procedure for which* $B_{\text{sam}} = 1000$ *random permutations were used. Note that the results for the Bonferroni, Holm and Hochberg procedures are virtually identical, similarly for the unadjusted p-value (rawp) and SAM Tusher procedures.*

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). Let x_{ji} denote the base 2 logarithm of the Cy5/Cy3 fluorescence intensity ratio for probe j in array i, $i = 1, \ldots, 16$, $j = 1, \ldots, 6{,}356$. Then x_{ji} represents the expression response of gene j in either a control ($i = 1, \ldots, 8$) or a treatment ($i = 9, \ldots, 16$) mouse.

Differentially expressed genes were identified by computing two-sample Welch t-statistics for each gene j,

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

where $\bar{x}_{1j}$ and $\bar{x}_{2j}$ denote the average expression measure of gene j in the $n_1 = 8$ control and $n_2 = 8$ treatment hybridizations, respectively. Similarly, s_{1j}^2 and s_{2j}^2 denote the variance of gene j's expression measure 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. 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.

3.1.2 *Bacteria experiment of 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 treatment with pathogenic bacterial components. The ability of an organism to combat microbial infection is crucial to survival. Humans, along with other higher organisms, possess two parallel, yet interacting, systems of defense against microbial invasion, referred to as the innate and adaptive immune systems. It has recently been discovered that cells of the innate immune system possess receptors which enable them to differentiate between the cellular components of different pathogens, including Gram-positive and Gram-negative bacteria, which differ in their cell wall structure, fungi, and viruses. Boldrick et al. sought to determine if, given the presence of specific receptors, cells of the innate immune system would have differing genomic responses to diverse pathogenic components. One important question that was addressed involved the response of these innate immune cells (PBMCs) to different doses of bacterial components. Although one can experimentally treat cells with the same amount of two types of bacteria, because the bacteria may differ in size or composition, one cannot be sure that this nominally equivalent treatment is truly equivalent, in the sense that the *strength* of the stimulation is equivalent. To make a statement that the response of PBMCs to a certain bacterium is truly specific to that bacterium, one must therefore perform a dose–response analysis to ensure that one is not simply sampling from two different points on the same dose–response curve.

Boldrick et al. performed a set of experiments (dose–response data set) that monitored the effect of three factors on the expression response of PBMCs: bacteria type, dose of the bacterial component and time after treatment. Two types of bacteria were considered: the Gram-negative *B. pertussis* and the Gram-positive *S. aureus*. Four doses of the pathogenic components were administered based on a standard dose: $1X$, $10X$, $100X$, $1000X$, where X represents the number of bacterial particles per human cell ($X = 0.002$ for the Gram positive and $X = 0.004$ for the Gram negative). The gene expression response was measured at five time points after treatment: 0.5, 2, 4, 6 and 12 hours (extra time points 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 immune cells or which are of known immunologic 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 six 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. (2002). For each microarray i, $i = 1, \ldots, 44$, 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 Gram-negative bacteria for one of the 22 dose $\times$ time combinations (4 doses $\times$ 5 time points plus 2 extra time points for dose $100X$). The analysis below is based on a subset of 2,562 genes that were well measured in both the dose–response and the diversity data sets (see Boldrick et al., 2002, 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

treatment by Gram-positive and Gram-negative bacterial components. 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_h denote the difference in the expression response to infection by the Gram-negative and Gram-positive bacteria for the hth dose $\times$ time block, $h = 1, \ldots, 22$. The paired t-statistic is defined as $t = \bar{d}/\sqrt{s_d^2/n_d}$, where $\bar{d}$ is the average of the $n_d = 22$ differences d_h and s_d^2 is the variance of these 22 differences. 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 the expression profiles *within* the 22 dose $\times$ time blocks (i.e., all permutations of the Gram-positive and Gram-negative labels within dose $\times$ time blocks).

3.1.3 *Leukemia study of Golub et al.* Correct diagnosis of neoplasia is necessary for proper treatment. The traditional means of identification and classification of malignancies has been based upon histology and immunohistologic staining of pathologic specimens. However, it is apparent, based upon the variability of response to treatment and length of survival after therapy, that there is variability within the current system of classification. Genomic technologies may provide the means by which neoplasms can be more accurately characterized and classified, thus leading to more effective diagnosis and treatment.

As a demonstration of such capabilities, Golub et al. (1999) studied two hematologic malignancies: acute lymphoblastic leukemia (ALL, class 1) and acute myeloid leukemia (AML, class 2), which are readily classifiable by traditional pathologic methods. They sought to show that these two malignant entities could be identified and distinguished based on microarray gene expression measures alone. Therefore, one of the goals of the statistical analysis was to identify genes that differed most significantly between the two diseases. 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 www.genome.wi.mit.edu/MPR). Following Golub et al. (Pablo Tamayo, personal communication), three preprocessing steps were applied to the normalized matrix of intensity values available on the website: (1) thresholding—floor of 100 and ceiling of 16,000; (2) filtering—exclusion of genes with max / min ≤ 5 or (max $-$ min) ≤ 500, where max and min refer, respectively, to the maximum and minimum intensities for a particular gene across mRNA samples; (3) base 10 logarithmic transformation. Box plots of the expression measures for each of the 38 samples revealed the need to standardize the expression measures 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 measure for gene j in mRNA sample i.

Differentially expressed genes in ALL and AML patients were identified by computing two-sample Welch t-statistics for each gene j as in Section 3.1.1. 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.

3.2 Simulated Data

While it is informative to examine the behavior of different multiple testing procedures on the microarray data sets described above, the genes cannot be unambiguously classified as differentially expressed or not. As a result, there is no "gold standard" for assessing Type I and Type II errors. Simulation studies are thus needed to evaluate the Type I error rate and power properties of each of the procedures. Artificial gene expression profiles $\mathbf{x}$ and binary responses y were generated as in Box 3 for $m = 500$ genes.

The 17 multiple testing procedures described in Table 3 were applied to each of the simulated data sets. Unadjusted p-values for each of the genes were computed in two ways: by permutation of the $n = n_1 + n_2$ responses and from the t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Table 4 lists the different parameters used in the simulation.

4. RESULTS

4.1 Microarray Data

The multiple testing procedures of Section 2 were applied to the three microarray data sets described in Section 3.1, using permutation to estimate unadjusted and adjusted p-values. For a given procedure, genes with permutation adjusted p-values $\tilde{p}_j^* \leq \alpha$ were declared to be differentially expressed at nominal level α for the Type I error rate controlled by the procedure under consideration. For each data set, ordered adjusted p-values were plotted for each procedure in panels (a)

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. Model parameters used in the simulation are listed in Table 4.
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 measures 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. A summary of the multiple testing procedures applied in the simulation study is given in Table 3.
3. For each procedure, record the number R_b of genes declared to be differentially expressed, the numbers V_b and T_b of Type I and II errors, respectively, and the false discovery rate Q_b, where $Q_b = V_b/R_b$ if $R_b > 0$ and $Q_b = 0$ 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:

$$\text{PCER} = \frac{\sum_b V_b/m}{B},$$

$$\text{FWER} = \frac{\sum_b I(V_b \geq 1)}{B},$$

$$\text{FDR} = \frac{\sum_b Q_b}{B},$$

$$\text{Average power} = 1 - \frac{\sum_b T_b/(m - m_0)}{B}.$$

and (b) of Figures 3–5 (see Section 2.8 for guidelines on interpreting the figures). Different line types are used for different Type I error rate definitions: solid, dashed, and dotted lines are used for FWER, FDR and PCER controlling procedures, respectively. Panels (c) of Figures 3–5 display plots of adjusted p-values (on a log scale) versus t-statistics. Finally, panels (d) of Figures 3–5 display permutation 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). As expected, for a given nominal α-level, the number of genes declared to be differentially expressed was the greatest for procedures controlling the PCER and the smallest for procedures that control the FWER. Indeed, adjusted p-values are the smallest for procedures that control the PCER (dotted curves in panels (a) and (b) of Figures 3–5 for SAM Tusher and for the procedure based on unadjusted p-values, rawp) and the largest for procedures that control the FWER (solid curves for Bonferroni, Holm, Hochberg and maxT procedures). Also as expected, the SAM Tusher procedure and the standard unadjusted p-value procedure (rawp) for controlling the PCER produced very similar results (overlap of the brown and goldenrod dotted curves for rawp and SAM Tusher in panels (a) and (b) of the three figures). As in the simulation study, the Benjamini and Yekutieli (2001) FDR procedure was much more conservative than the standard Benjamini and Hochberg (1995) procedure (magenta and cyan dashed curves in panels (a) and (b) of Figures 3–5 for BY and BH, respectively). For control of the FWER, procedures based on the step-down $\max T$ 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 (solid curves in panels (a) and (b) of Figures 3–5 for the four procedures that control the FWER).

The different multiple testing procedures behaved similarly for the leukemia and bacteria data sets; however, their behavior on the apo AI data set was quite different due to the smaller sample sizes. Aside from the PCER procedures, only the $\max T$ and standard Benjamini and Hochberg (1995) procedures rejected any hypothesis at nominal 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 data set clearly highlights the power of the $\max T$ procedure over standard Bonferroni-like procedures or even some procedures that control the FDR.

Apo AI experiment. In this experiment, eight spotted probe sequences clearly stood out from the remaining sequences: they had the largest absolute t-statistics and the smallest adjusted p-values for all procedures (see

TABLE 3
Multiple testing procedures applied in the 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 max T procedure, reject H_j if $\tilde{p}^*_j \leq \alpha$ [Equation (4)]
maxT sd	Step-down max T procedure, reject H_{r_j} if $\tilde{p}^*_{r_j} \leq \alpha$ [Equation (8), Box 2]
FDR BH t	Benjamini and 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 and 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 and 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 and 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 tusher	Tusher, Tibshirani and Chu (2001) SAM procedure (Section 2.7.2), reject $H_{(j)}$ if $\tilde{p}^*_{(j)} \leq \alpha$, estimated by permutation
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
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

Note. The reader is referred to the technical report by Dudoit, Shaffer and Boldrick (2002) for details on adjusted p-value calculations for SAM and for the step-down and step-up versions of Golub et al.'s (1999) neighborhood analysis.

TABLE 4
Simulation parameters

Parameter	Simulation A	Simulation B	Simulation C	Simulation D
Number of genes, m	500	500	500	500
Mean vectors				
μ_1	0_m	0_m	0_m	0_m
μ_2	0_m	0_m	$[b_{m\cdot 0.1}, -b_{m\cdot 0.1}, 0_{m\cdot 0.8}]$	$[b_{m\cdot 0.1}, -b_{m\cdot 0.1}, 0_{m\cdot 0.8}]$
Covariance matrix, Σ	S_m	S_m	S_m	S_m
Sample sizes				
n_1	25	5	25	5
n_2	25	5	25	5
Number of simulations, B	500	500	500	500
Number of permutations for SAM, B_{sam}	1000	$\binom{n_1+n_2}{n_1}$	1000	$\binom{n_1+n_2}{n_1}$
Number of permutations for neighborhood analysis, B_{nbd}	1000	$\binom{n_1+n_2}{n_1}$	1000	$\binom{n_1+n_2}{n_1}$
Number of permutations for unadjusted p-values, B_{perm}	25,000	$\binom{n_1+n_2}{n_1}$	25,000	$\binom{n_1+n_2}{n_1}$
Nominal Type I error rate, α (PCER, FWER or FDR)	0.05	0.05	0.05	0.05

Note. Here, 0_n denotes an n-vector with entries equal to 0 and b_n denotes the n-vector $1.5 \cdot (1, 2, \ldots, n)/n$. S_m is the $m \times m$ covariance matrix for a random subset of m genes in the Boldrick et al. (2002) experiment described in Section 3.1.2.

drop for the smallest eight t-statistics in the Q–Q plot of Figure 3, panel (d)). In particular, all eight $\max T$ adjusted p-values were less than 0.05. The negative t-statistics suggest that the genes are under-expressed in the apo AI knock out mice compared to the control mice. The eight probe sequences actually correspond to only four distinct genes: apo AI (three copies), apo CIII (two copies), sterol C5 desaturase (two copies), and a novel EST (one 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, because 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. (2000) 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 $\max T$ adjusted p-values less than 0.05 and several of these sequences actually represented different clones of the same genes: CD64 (three copies), Iκ B alpha (five copies), SHP-1 (two copies) and plasma gelsolin (two copies) (see gene list in Appendix A of Dudoit, Shaffer and Boldrick (2002)). In contrast to the apo AI experiment, the genes exhibited a continuum of change in expression and we could not identify a group of genes that clearly stood out from the rest. This is likely due in part to the biological nature of the experiment, in which the two bacterial treatments were very similar in their stimulatory effect and extreme differences in gene expression are not present. As discussed in Section 2.8 and illustrated in panel (c) of Figure 4, the multiple testing procedures fall into two main categories: those for which adjusted p-values are monotone in the test statistics (maxT and SAM Tusher, i.e., purple and goldenrod plotting symbols in panel (c), respectively), and those for which adjusted p-values are monotone in the unadjusted p-values (all other procedures). Within each of these classes, the procedures produce the same gene orderings and differ only in the cutoffs for significance. Figure 6 displays a comparison of the gene orderings based on absolute t-statistics, $|t_j|$, and permutation unadjusted p-values, p_j^*. It is a plot, for each number of genes G, of the proportion of genes having both the G largest absolute t-statistics and the G smallest permutation unadjusted p-values, that is, a plot of $|\{1 \le j \le m : p_j^* \le p_{(G)}^* \text{ and } |t_j| \ge |t|_{(G)}\}|/G$ versus G. There are some discrepancies between the two orderings, especially among the 10 most significant genes found by each criterion. The overlap proportion can be as low as 25% for $G = 4$; for $G = 100$ onward, the agreement exceeds 80%. Discrepancies arise because the test statistics T_j of different genes have different permutation distributions. A detailed discussion of the biological findings can be found in Boldrick et al. (2002).

Leukemia study. For this data set, 92 genes had $\max T$ adjusted p-values less than 0.05 [see gene list in Appendix B of Dudoit, Shaffer and Boldrick (2002)]. There is significant overlap between this list and the gene list in Golub et al. (1999, page 533 and Figure 3B). Refer to Golub et al. for a description of the genes and their involvement in ALL and AML. Additional figures and detailed discussions of the results for SAM and neighborhood analysis are given in the technical report by Dudoit, Shaffer and Boldrick (2002). As with the bacteria experiment, the genes exhibited a continuum of change in expression and we could not identify a group of genes that stood out from the rest.

Several biological factors likely underlie the disparity between the three data sets. Whereas the apo AI experiment compares a relatively pure cell population (hepatocytes) from wild-type and knock out mice with an otherwise identical genetic background, the bacteria experiment and the leukemia study focus on comparisons of samples composed of a variety of cell types from genetically diverse individuals. In both of the latter cases, because of the nature of the complex samples studied, RNA may have been isolated from cell populations that were not affected by the variables being compared. For instance, within the leukemia study, one would anticipate that genes exhibiting strong differences between AML and ALL specimens would be myeloid- or lymphoid-specific, respectively. As the specimens studied invariably include some non-leukemic (i.e., normal) cells of myeloid or lymphoid origin, the contribution of gene expression from these cells likely buffers the tumor-specific expression signatures. In the bacterial study, the dynamic nature of the experimental design (i.e., with variation in time and dose of bacteria), and the presence of cell types unaffected by the bacterial treatment, similarly confounds such analysis.

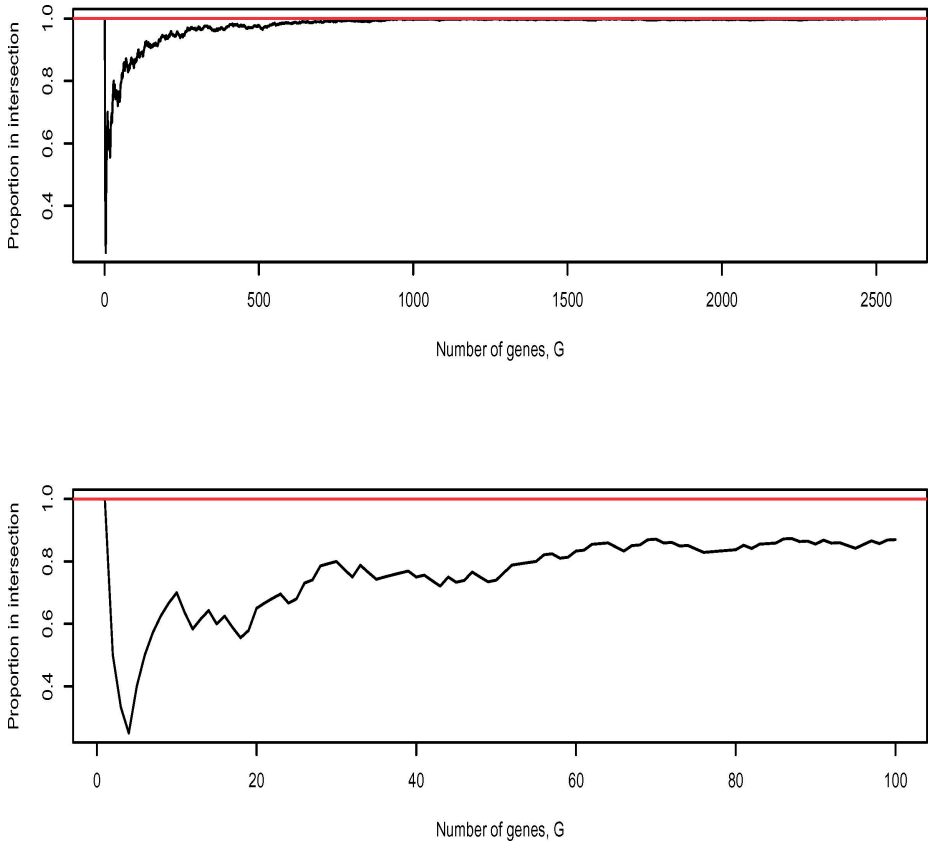

FIG. 6. *Bacteria experiment. The proportion of genes having both the G largest absolute t-statistics and the G smallest permutation unadjusted p-values is plotted versus G: $|\{1 \le j \le m : p_j^* \le p_{(G)}^*$ and $|t_j| \ge |t|_{(G)}\}|/G$ versus G. The bottom panel is an enlargement of the top panel for $G \le 100$. The plots provide a comparison of the gene lists produced by the two main types of procedures described in Section* 2.8.

4.2 Simulated Data

Figures 7–9 display plots of Type I error rates and power for different multiple testing procedures in the simulation study (see Box 3 and Tables 3 and 4 for a description of the procedures and parameters for simulation models A–D). For each procedure, adjusted p-values were computed as detailed in Section 2, using both a t-distribution and permutation to obtain unadjusted p-values. Null hypotheses were rejected whenever the corresponding adjusted p-values were less than a prespecified level α. Procedures designed to control the FWER, FDR and PCER are labeled in purple, green and orange, respectively. For each definition of Type I error rate, red plotting symbols are used for the procedures which are supposed to control this error rate. With the exception of Golub sd and Golub su, all procedures controlled the claimed Type I error rate in the strong sense (see, for example, the PCER panels in Figures 8 and 9, where the Golub sd and Golub su actual PCER are much greater than the nominal 0.05 level). As expected, procedures that control the FWER were the most conservative, followed by procedures that control the FDR (power comparison in the bottom right panels of Figures 8 and 9).

Procedures that control the FWER. The simulation study allowed us to compare the performance of single-step versus stepwise procedures (i.e., Bonferroni versus Holm and Hochberg procedures, and single-step $\max T$ versus step-down $\max T$ procedures). Although stepwise procedures are generally less conservative than single-step procedures, we found that the difference was minute in our applications. This is to be expected in testing problems with a large number of null hypotheses m, most of which are true. In such cases, the correction $(m - k + 1)$ used in the Holm and Hochberg procedures is very close to the Bonferroni correction m for moderate k [see Equations (5) and (9), where k refers to the k hypotheses with the smallest unadjusted p-values]. In contrast, incorporating the dependence structure among the genes, as in

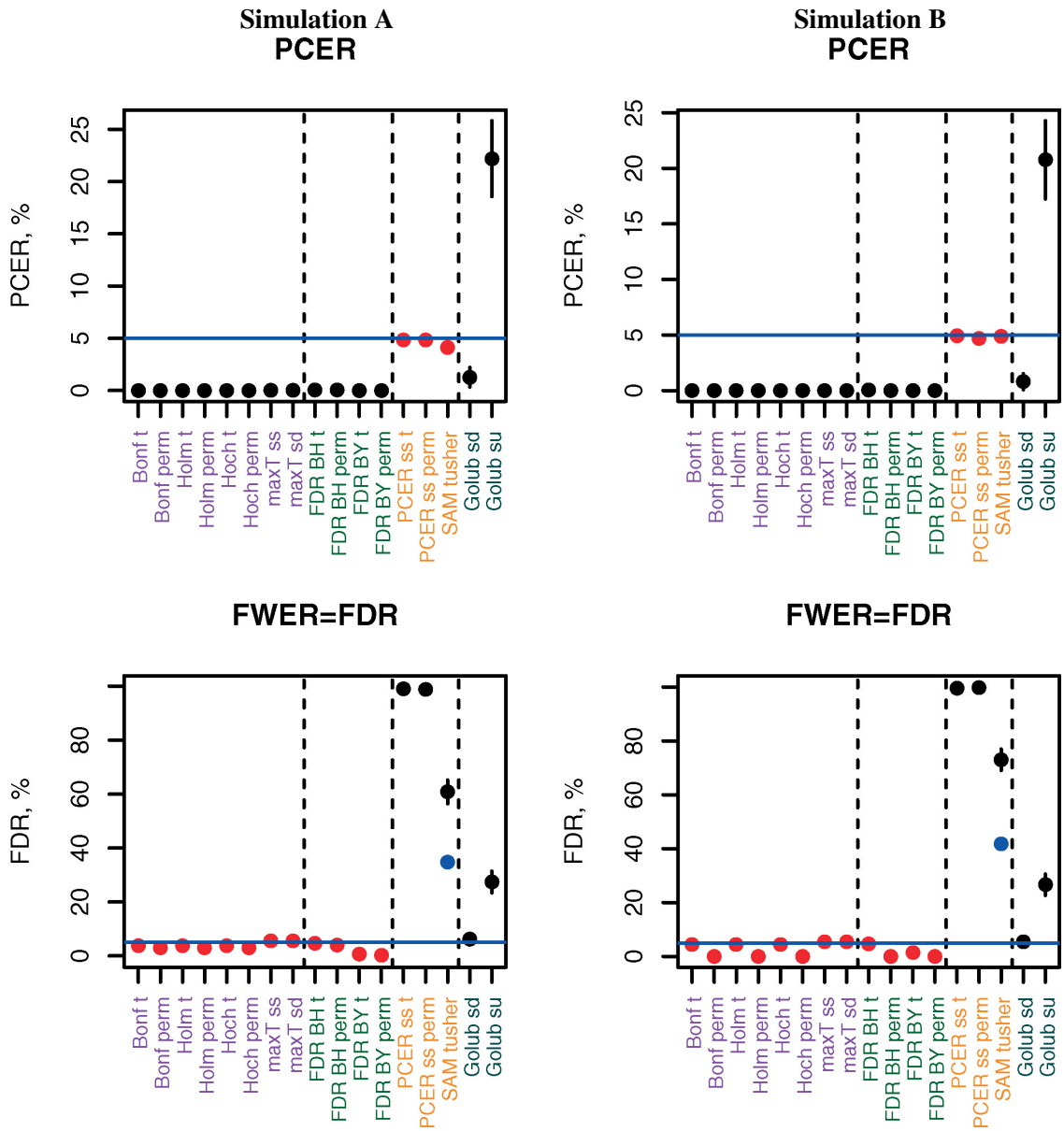

FIG. 7. *Simulations A and B—complete null.* PCER, FWER *and* FDR *for different multiple testing procedures in Simulation A (left) and Simulation B (right). The top panels display* $\text{PCER} = \sum_b R_b/mB$ *and simulation standard errors* (2 SE)*; the bottom panels display* $\text{FWER} = \text{FDR} = \sum_b I(R_b \geq 1)/B$ *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 nominal Type I error rate of* $\alpha = 5\%$*. In the FDR panels, the simulation average of the nominal SAM FDR is plotted in blue. Details on each of the multiple testing procedures and simulation parameters are given in Tables* 3 *and* 4*, respectively.*

the max T 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 sample sizes when the unadjusted p-values used in the Bonferroni, Holm and Hochberg procedures were estimated by permutation (for example, in the bottom right panel of Figure 9 for simulation model D, Bonf perm, Holm perm and Hoch perm have power around 0, while maxT ss and maxT sd have power around 8%).

Procedures that control 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 that control the PCER (for example, in the bottom right panel of Fig-

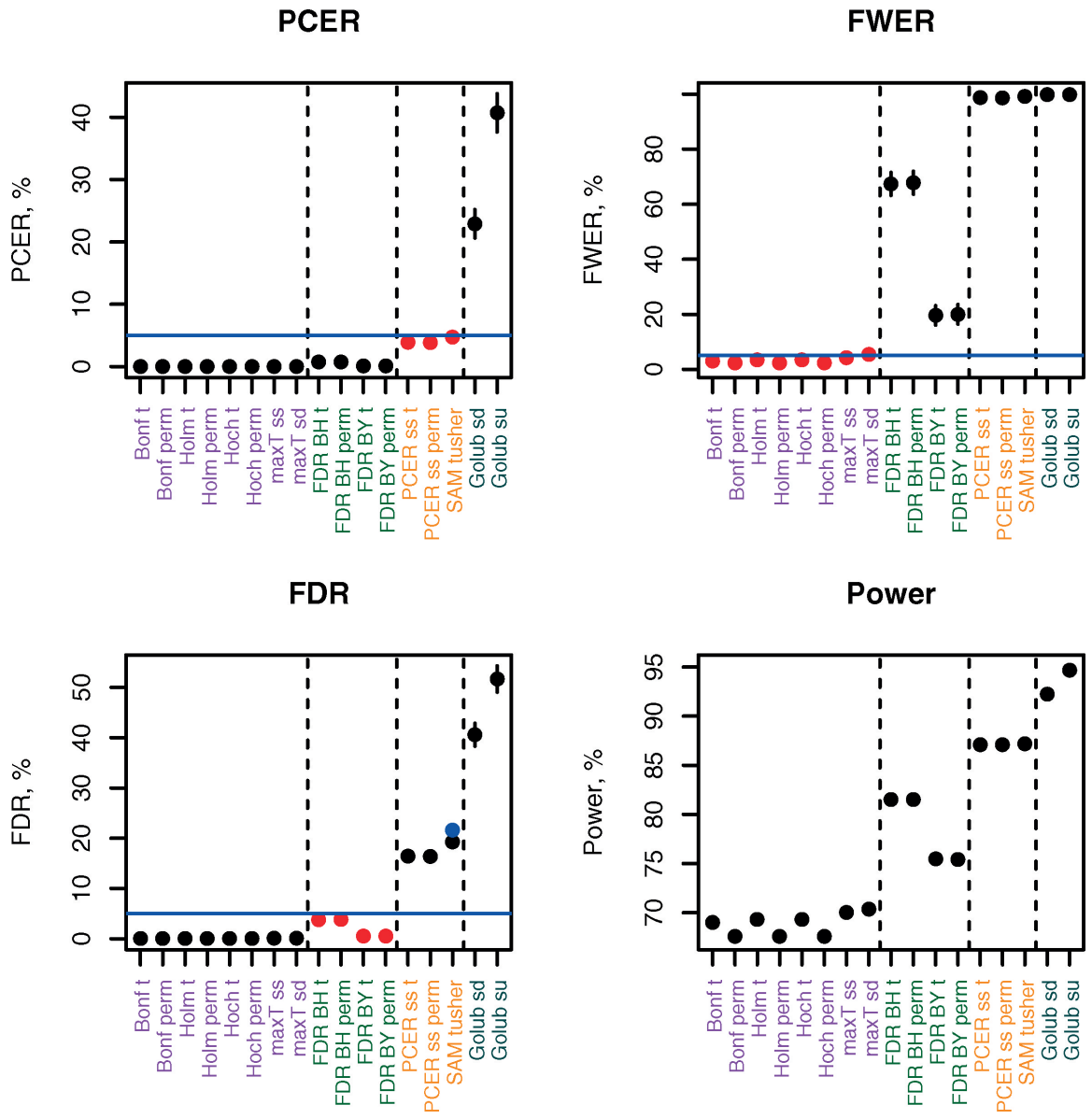

FIG. 8. *Simulation C—20% false nulls* ($m_0/m = 0.8$). PCER, FWER, FDR *and average power for different multiple testing procedures. The top left panel displays* PCER $= \sum_b V_b/mB$ *and simulation standard errors* (2 SE)*; the top right panel displays* FWER $= \sum_b I(V_b \geq 1)/B$ *and simulation 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 $= 1 - \sum_b T_b/(m - m_0)B$ *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 nominal Type I error rate of* $\alpha = 5\%$*. In the* FDR *panel, the simulation average of the nominal* SAM FDR *is plotted in blue. Details on each of the multiple testing procedures and simulation parameters are given in Tables* 3 *and* 4*, respectively.*

ure 8, the power of FDR BH perm is about 81% compared to about 70% for maxT sd and about 87% for PCER ss perm). Also as expected, the Benjamini and Yekutieli (2001) FDR procedure was more conservative than the Benjamini and Hochberg (1995) procedure (up to a 30% difference in power in Figure 9 for FDR BH t and FDR BY t) 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 and Hochberg procedure controlled the FDR at the nominal 5% level, in spite of the correlations among the test statistics.

SAM procedures. A detailed discussion and comparison of the SAM procedures in Efron et al. (2000) and Tusher, Tibshirani and Chu (2001), including the derivation of adjusted p-values, are found in Dudoit,

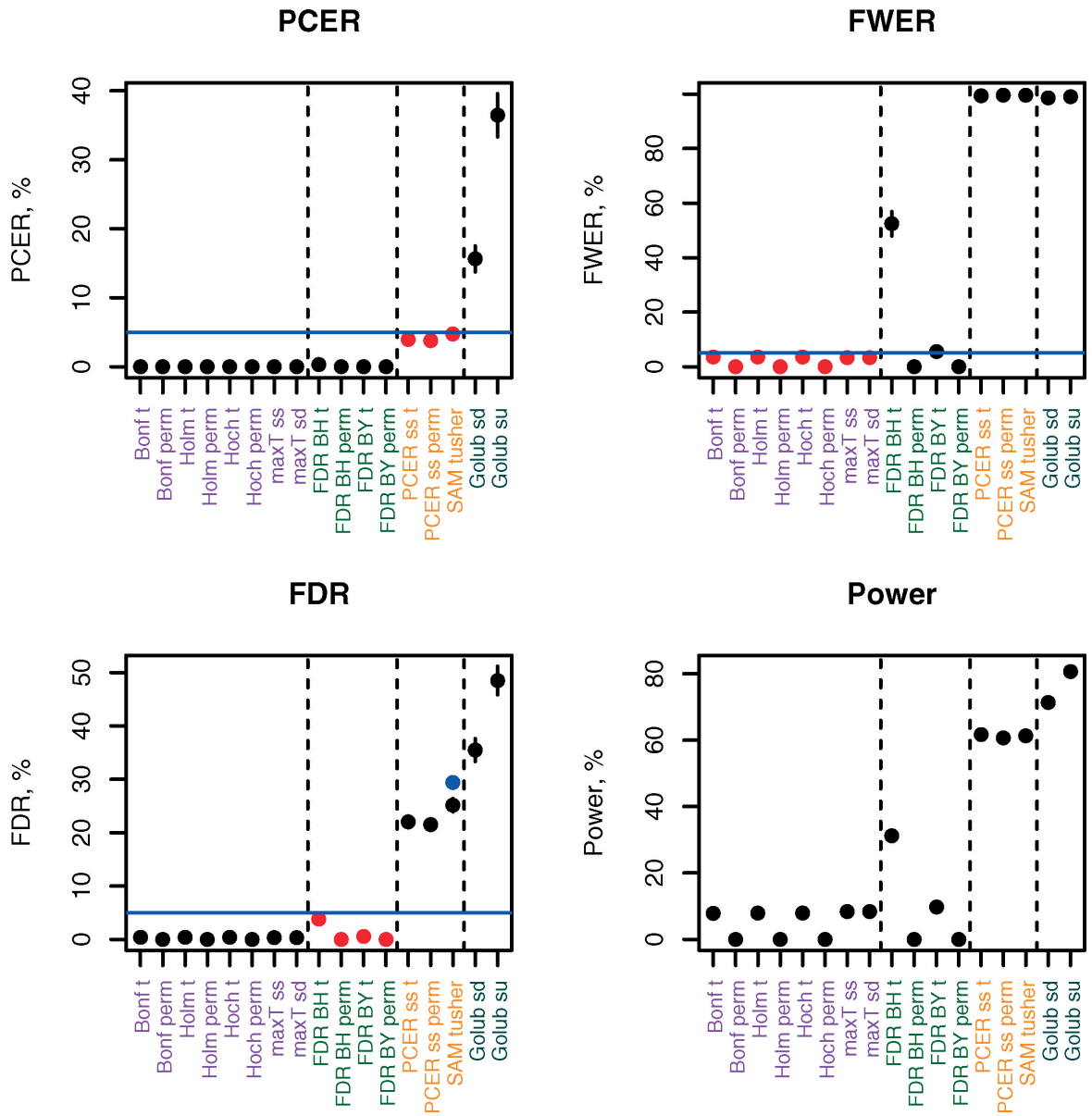

FIG. 9. *Simulation D—20% false nulls* ($m_0/m = 0.8$). PCER, FWER, FDR *and average power for different multiple testing procedures. The top left panel displays* PCER $= \sum_b V_b/mB$ *and simulation standard errors* (2 SE); *the top right panel displays* FWER $= \sum_b I(V_b \geq 1)/B$ *and simulation 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 $= 1 - \sum_b T_b/(m - m_0)B$ *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 nominal Type I error rate of* $\alpha = 5\%$. *In the FDR panel, the simulation average of the nominal SAM FDR is plotted in blue. Details on each of the multiple testing procedures and simulation parameters are given in Tables* 3 *and* 4*, respectively.*

Shaffer and Boldrick (2002). In contrast to the earlier version of SAM in Efron et al. (2000), the SAM procedure in Tusher, Tibshirani and Chu (2001) is not entirely based on the permutation distribution of the order statistics and controls the PCER in the strong sense. The FDR panels of Figures 7–9 display the average of the *nominal* SAM FDR ($\widehat{\text{FDR}}_b^0 = \widehat{\text{PFER}}_b^0/R_b$, where $\widehat{\text{PFER}}_b^0$ is the SAM estimate of the PFER for the bth simulation), 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 nominal SAM FDR was actually greater than 1. SAM is very similar in power to standard procedures that control the PCER in the strong sense (compare the power for SAM tusher

to the power for PCER ss t and PCER ss perm in the bottom right panels of Figures 8 and 9).

Neighborhood analysis. As shown in Figures 7–9, the step-down version of neighborhood analysis controls the FWER under the complete null (weak control), but fails to do so when there are false null hypotheses. The step-up version of neighborhood analysis does not control any known type of error rate, not even the PCER, and can lead to very high Type I error rates (see the PCER, FDR and FWER panels for Golub sd and Golub su in Figures 8 and 9). A detailed discussion of neighborhood analysis is given in Dudoit, Shaffer and Boldrick (2002).

Nominal t-distribution versus permutation p-values. Because the gene expression measures 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 results confirm that, in this case, 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 differences being for small sample sizes ($n_1 = n_2 = 5$, for Simulation B in Figure 7, right panel, and Simulation D in Figure 9). 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 step-up procedures and the Bonferroni, Holm and Hochberg procedures. Procedures based on the max T adjusted p-values, which involve the test statistics rather than the unadjusted p-values, did not suffer from this problem.

5. DISCUSSION

In this article, we have discussed different approaches to large-scale multiple hypothesis testing in the context of DNA microarray experiments. Standard multiple testing procedures, as well as recent and oft-cited proposals for microarray experiments, were compared in terms of their Type I error rate control and power properties, using actual gene expression data sets and by simulation.

The comparison study highlighted five desirable features of multiple testing procedures for large multiplicity problems such as those arising in microarray experiments: (1) control of an appropriate and precisely defined *Type I error rate*; (2) *strong control* of the Type I error rate, that is, control of this error rate under any combination of true and false null hypotheses corresponding to the true data generating distribution; (3) accounting for the *joint distribution* of the test statistics; (4) reporting the results in terms of *adjusted p-values*; (5) 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, Martin and Churchill, 2000; Manduchi et al., 2000; Pollard and van der Laan, 2003; Tusher, Tibshirani and Chu, 2001; Westfall, Zaykin and Young, 2001). However, not all proposed solutions were cast within a standard statistical framework and some fail to provide adequate Type I error rate control. In particular, the Type I error rates considered in some of these articles were rather loosely defined, thus making it difficult to assess the properties of the multiple testing procedures. Regarding item (1), control of the per-comparison error rate 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, Tibshirani and Chu (2001), both SAM procedures are based on control of 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 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 and in the technical report by Dudoit, Shaffer and Boldrick (2002). In the microarray setting, where it is very unlikely that no genes are differentially expressed, property (2) of strong control of the Type I error rate is essential, whether it be the FWER, PCER or FDR. The simulation study demonstrated that some of the procedures did not provide strong control of the Type I error rate, that is, the Type I error rate was no longer controlled at the nominal level when a subset of null hypotheses was 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 provide only *weak* control of the Type I error rate. Regarding point (3), the comparison study highlighted the gains in power that can be achieved by taking into account

the joint distribution of the test statistics when assessing statistical significance (max T procedures versus 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 4) to be particularly useful and flexible summaries of the strength of the evidence against the null. The adjusted p-value for a particular gene reflects the nominal false positive rate for the entire experiment when genes with smaller p-values are declared to be differentially expressed. Adjusted p-values may also be used to summarize and compare the results from different multiple testing procedures as described in Section 2.8. Finally, as mentioned in item 5, efficient resampling-based nonparametric multiple testing procedures are needed to take into account the complex and unknown dependence structures among the expression measures of different genes. Resampling procedures were proposed in Westfall and Young (1993) for FWER control; however, due to the large-scale nature of microarray experiments, computational issues remain to be addressed in addition to methodological ones (Ge, Dudoit and Speed, 2003).

Procedures that control the FWER. Results on both simulated and microarray data sets suggest that the Westfall and Young (1993) step-down max T procedure is well-adapted for DNA 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 among the test statistics of different genes. In addition, the max T procedure performed very well compared to other procedures (including some FDR procedures), when adjusted p-values were estimated by permutation. Because the max T adjusted p-values are based on the test statistics rather than the unadjusted p-values, the procedure does not suffer as much as others from the small number of possible permutations associated with small sample sizes.

Procedures that control the FDR. In the microarray setting, where thousands of tests are performed simultaneously and a fairly large number of genes are expected to be differentially expressed, procedures that control the FDR present a promising alternative to approaches that control 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 dependence structures (e.g., independence or positive regression dependence) or do not exploit the joint distribution of the test statistics. It would thus be useful to develop procedures, in the spirit of the Westfall and Young (1993) min P and max T 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 and Benjamini (1999), assuming unadjusted p-values for the true null hypotheses are independent of the p-values for the false null hypotheses. Reiner, Yekutieli and Benjamini (2001) applied different FDR controlling procedures to the apo AI data set.

SAM procedures. The Efron et al. (2000) and Tusher, Tibshirani and Chu (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, $E(V \mid \mathrm{H}_0^C)$, leading to the choice of the threshold Δ. The difference lies in the use of ordered test statistics in Efron et al. (2000) 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, Tibshirani and Chu (2001) version it is *strongly* controlled. The only difference between the latter 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, Tibshirani and Chu (2001) SAM procedure (software package 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. The use of adjusted p-values, derived in Dudoit, Shaffer and Boldrick (2002), provides a more flexible implementation of the procedure. As part of the SAM method, Efron et al. (2000) and Tusher, Tibshirani and Chu (2001) suggest test statistics for identifying differentially expressed genes for different types of responses and covariates. 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 and Speed, 2002). The "shrunken" statistics were not used in the comparison study in Section 4, because we wanted to focus on Type I error rate control 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 nominal level α. Dudoit, Shaffer and Boldrick (2002) considered a step-down and a step-up version of neighborhood analysis to deal with the nonmonotonicity of $G(c)$ and derived corresponding adjusted p-values. Because neighborhood analysis is based on the distribution of 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 neighborhood analysis to the three microarray data sets 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 to be differentially expressed. Golub et al. applied the neighborhood analysis separately for each type of one-sided hypothesis (overexpression in AML compared to ALL and vice versa); it is not clear how an overall Type I error rate can be obtained.

5.1 Open Questions

In many situations, DNA microarray experiments are used as a first exploratory step in the process of identifying subsets of genes involved in important biological processes. Genes identified by microarray analysis are typically followed up using other assays such as RT-PCR or in situ hybridization. In this exploratory context, one may be more interested in minimizing the Type II error (i.e., maximizing power) rather than minimizing the Type I error, that is, one may be willing to tolerate a larger number of false positives in order to capture as many "interesting" genes as possible. Contrary to common belief, multiple testing approaches are actually very relevant to such exploratory analyses. The reporting methods described in Section 2.8 can be used gainfully in an exploratory setting by allowing researchers to select an appropriate combination of number of genes and tolerable false positive rate for a particular experiment and available resources. Receiver Operating Characteristic (ROC) curves provide useful tools for examining Type I and Type II error properties (Pepe et al., 2003). While test optimality (in terms of power) is a well-established subject in univariate hypothesis testing (e.g., uniformly most powerful tests), very few optimality results are available in the multivariate setting (Hochberg and Tamhane, 1987). Given suitable definitions of Type I and Type II error rates, very little is known about procedures which minimize the Type II error rate for a given level of Type I error. Optimality of multiple tests is an interesting research avenue to pursue from both a theoretical and practical point of view.

Gene prescreening is a common issue in expression and other large-scale biological experiments. By reducing the number of tests, prescreening is often seen as a means of increasing power. In microarray experiments, preliminary screening of the genes is generally done based on data quality criteria, such as signal to background intensity and proportion of missing values. A natural question, then, is whether the Type I error rate is controlled at the claimed level. The answer depends on the screening criterion. Control is achieved in the following cases: (1) a gene subset is selected based on subject matter knowledge *before* looking at the data and (2) the statistics used for screening are independent of the test statistics under the null hypotheses. Other situations still need to be better understood.

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, that is, 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 incorrectly 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; Shaffer, 2002) and was not 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 3.1.2), 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 two-dimensional testing problem where several hypotheses are tested simultaneously for each of thousands of genes [see Gabriel (1975), Krishnaiah and Reising (1985), Morrison (1990), and Westfall and Young (1993) for background on multivariate multiple testing methods].

A related issue is the development of resampling methods for multiple testing in the context of factorial or time course experiments which impose some structure on the columns of the gene expression data matrix. For the three-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 that can be used to estimate unadjusted and adjusted p-values (Westfall and Young, 1993). Bootstrap procedures, both parametric and nonparametric, should also be investigated, because they allow consideration of more specific null hypotheses (for example, *equal mean* gene expression levels for two types of cell populations, as opposed to the stronger null hypothesis of *identical distributions* imposed by permutation procedures) and may lead to increased power (Pollard and van der Laan, 2003).

The methods described above concern hypotheses about 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. (2002). Gene subset selection procedures based on resampling are described in van der Laan and Bryan (2001).

Other approaches. The present article focused on frequentist approaches to multiple testing. In particular, we did not consider Bayesian procedures, which constitute an important class of methods for the identification of differentially expressed genes and whose thorough treatment would require a separate article. Applications of Bayesian procedures in microarray data analysis can be found in Efron, Storey and Tibshirani (2001), Efron et al. (2001), Manduchi et al. (2000), and Newton et al. (2001). In such methods, the criterion for identifying differentially expressed genes is based on the posterior probability of differential expression, that is, the probability that a particular gene is differentially expressed *given* the data for all genes. This is in contrast to frequentist methods, which are based on adjusted p-values, that is, 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) discussed Bayesian interpretations of the FDR. An interesting philosophical approach to statistical inference is found in the recent work of Mayo and Spanos (2002). These authors provided a post-data or posterior interpretation of frequentist tests based on a severity principle.

SOFTWARE

Most of the multiple testing procedures considered in this article were implemented in an R package (Ihaka and Gentleman, 1996), multtest, which may be downloaded from http://www.bioconductor.org. The package includes procedures for controlling the family-wise error rate (FWER): Bonferroni, Hochberg (1988), Holm (1979), Šidák, Westfall and Young (1993) min P and max T. It also includes procedures for controlling the false discovery rate (FDR): Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) step-up procedures. The procedures are implemented for tests based on t- and F-statistics for one- and two-factor designs. Permutation procedures are available to estimate unadjusted and adjusted p-values. The website http://www.math.tau.ac.il/∼roee/index.htm provides references and software related to FDR controlling procedures.

Note Added in Proof. A previous version of this article contained a review of an earlier SAM procedure that appeared in the technical report by Efron et al. (2000). However, these authors no longer endorse that procedure and it was decided to eliminate it from the final version of the present article. A detailed discussion and comparison of both SAM procedures is given in the technical report by Dudoit, Shaffer and Boldrick (2002).

ACKNOWLEDGMENTS

The authors are most grateful to Warren Ewens, Yongchao Ge, Gregory Grant, Mark van der Laan,

Erich Lehmann and Peter Westfall for insightful discussions on multiple testing. They would also like to acknowledge the editors and two referees for their constructive comments on an earlier version of the article.

REFERENCES

ALIZADEH, A. A., EISEN, M. B., DAVIS, R. E., MA, C., LOSSOS, I. S., ROSENWALD, A., BOLDRICK, J. C., SABET, H., TRAN, T., YU, X., POWELL, J. I., YANG, L., MARTI, G. E., MOORE, T., HUDSON JR., J., LU, L., LEWIS, D. B., TIBSHIRANI, R., SHERLOCK, G., CHAN, W. C., GREINER, T. C., WEISENBURGER, D. D., ARMITAGE, J. O., WARNKE, R., LEVY, R., WILSON, W., GREVER, M. R., BYRD, J. C., BOTSTEIN, D., BROWN, P. O. and STAUDT, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403** 503–511.

ALON, U., BARKAI, N., NOTTERMAN, D. A., GISH, K., YBARRA, S., MACK, D. and LEVINE, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.* **96** 6745–6750.

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

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188.

BERAN, R. (1988). Balanced simultaneous confidence sets. *J. Amer. Statist. Assoc.* **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. and RELMAN, D. A. (2002). Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **99** 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. and BOTSTEIN, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21** 33–37.

BUCKLEY, M. J. (2000). *The Spot User's Guide*. CSIRO Mathematical and Information Sciences, North Ryde, NSW, Australia. Available at http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm.

CALLOW, M. J., DUDOIT, S., GONG, E. L., SPEED, T. P. and RUBIN, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research* **10** 2022–2029.

CHU, G., GOSS, V., NARASIMHAN, B. and TIBSHIRANI, R. (2000). SAM (Significance Analysis of Microarrays)—Users guide and technical document. Technical report, Stanford Univ.

DUDOIT, S., SHAFFER, J. P. and BOLDRICK, J. C. (2002). Multiple hypothesis testing in microarray experiments. Technical Report 110, Division of Biostatistics, Univ. California, Berkeley. Available at http://www.bepress.com/ucbbiostat/paper110/.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. and SPEED, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica* **12** 111–139.

DUNN, O. J. (1958). Estimation of the means of dependent variables. *Ann. Math. Statist.* **29** 1095–1111.

EFRON, B., STOREY, J. D. and TIBSHIRANI, R. (2001). Microarrays, empirical Bayes methods, and false discovery rates. Technical Report 2001-23B/217, Dept. Statistics, Stanford Univ.

EFRON, B., TIBSHIRANI, R., GOSS, V. and CHU, G. (2000). Microarrays and their use in a comparative experiment. Technical Report 2000-37B/213, Dept. Statistics, Stanford Univ.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160.

FINNER, H. (1999). Stepwise multiple test procedures and control of directional errors. *Ann. Statist.* **27** 274–289.

GABRIEL, K. R. (1975). A comparison of some methods of simultaneous inference in MANOVA. In *Multivariate Statistical Methods: Among-Groups Covariation* (W. R. Atchley and E. H. Bryant, eds.) 61–80. Dowden, Hutchinson and Ross, Stroudsburg, PA.

GE, Y., DUDOIT, S. and SPEED, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *TEST*. To appear.

GENOVESE, C. and WASSERMAN, L. (2001). Operating characteristics and extensions of the FDR procedure. Technical Report 737, Dept. Statistics, Carnegie Mellon Univ.

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. and 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.

HOCHBERG, Y. and TAMHANE, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

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. and BERNHARD, G. (1999). Bonferroni procedures for logically related hypotheses. *J. Statist. Plann. Inference* **82** 119–128.

IHAKA, R. and GENTLEMAN, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5** 299–314.

JOGDEO, K. (1977). Association and probability inequalities. *Ann. Statist.* **5** 495–504.

KERR, M. K., MARTIN, M. and CHURCHILL, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7** 819–837.

KRISHNAIAH, P. R. and REISING, J. M. (1985). Multivariate multiple comparisons. *Encyclopedia of Statistical Sciences* **6** 88–95. Wiley, New York.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

LIPSHUTZ, R. J., FODOR, S., GINGERAS, T. R. and LOCKHART, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics* **21** 20–24.

LÖNNSTEDT, I. and SPEED, T. P. (2002). Replicated microarray data. *Statist. Sinica* **12** 31–46.

MANDUCHI, E., GRANT, G. R., MCKENZIE, S. E., OVERTON, G. C., SURREY, S. and STOECKERT JR., C. J. (2000). Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics* **16** 685–698.

MAYO, D. and SPANOS, A. (2002). A severe testing interpretation of Neyman–Pearson tests. Unpublished.

MORRISON, D. F. (1990). *Multivariate Statistical Methods*, 3rd ed. McGraw-Hill, New York.

NATIONAL READING PANEL (1999). Teaching children to read. Report, National Institute of Child Health and Human Development, National Institutes of Health.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. and 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.

PEPE, M. S., LONGTON, G., ANDERSON, G. L. and SCHUMMER, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**. To appear.

PEROU, C. M., JEFFREY, S. S., VAN DE RIJN, M., REES, C. A., EISEN, M. B., ROSS, D. T., PERGAMENSCHIKOV, A., WILLIAMS, C. F., ZHU, S. X., LEE, J. C. F., LASHKARI, D., SHALON, D., BROWN, P. O. and BOTSTEIN, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *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. and BROWN, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* **23** 41–46.

POLLARD, K. S. and VAN DER LAAN, M. J. (2003). Resampling-based multiple testing with asymptotic strong control of type I error. Submitted.

RAMSEY, P. H. (1978). Power differences between pairwise multiple comparisons. *J. Amer. Statist. Assoc.* **73** 479–485.

REINER, A., YEKUTIELI, D. and BENJAMINI, Y. (2001). Using resampling-based FDR controlling multiple test procedures for analyzing microarray gene expression data. Unpublished.

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., REES, C., SPELLMAN, P., IYER, V., JEFFREY, S. S., VAN DE RIJN, M., WALTHAM, M., PERGAMENSCHIKOV, A., LEE, J. C. F., LASHKARI, D., SHALON, D., MYERS, T. G., WEINSTEIN, J. N., BOTSTEIN, D. and BROWN, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24** 227–234.

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York.

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

SHAFFER, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81** 826–831.

SHAFFER, J. P. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology* **46** 561–584.

SHAFFER, J. P. (2002). Multiplicity, directional (Type III) errors, and the null hypothesis. *Psychological Methods* **7** 356–369.

ŠIDÁK, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **62** 626–633.

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. *J. Amer. Statist. Assoc.* **84** 608–610.

STOREY, J. D. (2001). The false discovery rate: A Bayesian interpretation and the q-value. Technical Report 2001-12, Dept. Statistics, Stanford Univ.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498.

STOREY, J. D. and TIBSHIRANI, R. (2001). Estimating the positive false discovery rate under dependence, with applications to DNA microarrays. Technical Report 2001-28, Dept. Statistics, Stanford Univ.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B., EISEN, M., SHERLOCK, G., BROWN, P. and BOTSTEIN, D. (2002). Exploratory screening of genes and clusters from microarray experiments. *Statist. Sinica* **12** 47–59.

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

TUSHER, V. G., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **98** 5116–5121.

VAN DER LAAN, M. J. and BRYAN, J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics* **2** 445–461.

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

WESTFALL, P. H., ZAYKIN, D. V. and YOUNG, S. S. (2001). Multiple tests for genetic effects in association studies. In *Biostatistical Methods* (S. Looney, ed.) 143–168. Humana, Totowa, NJ.

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

YANG, Y. H., BUCKLEY, M. J., DUDOIT, S. and SPEED, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph. Statist.* **11** 108–136.

YANG, Y. H., DUDOIT, S., LUU, P. and SPEED, T. P. (2001). Normalization for cDNA microarray data. In *Microarrays: Optical Technologies and Informatics* (M. L. Bittner, Y. Chen, A. N. Dorsel and E. R. Dougherty, eds.) 141–152. SPIE, Bellingham, WA.

YEKUTIELI, D. and BENJAMINI, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference* **82** 171–196.