Published in final edited form as:

Stat Methods Med Res. 2009 December ; 18(6): 543–563. doi:10.1177/0962280209351899.

Multiple testing and its applications to microarrays

Yongchao Ge¹, Stuart C. Sealfon¹, and Terence P. Speed^{2,3}

¹Department of Neurology and Center for Translational Systems Biology Mount Sinai School of Medicine, New York

²Department of Statistics, University of California, Berkeley

³Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia

Abstract

The large-scale multiple testing problems resulting from the measurement of thousands of genes in microarray experiments have received increasing interest during the past several years. This paper describes some commonly used criteria for controlling false positive errors, including familywise error rates, false discovery rates and false discovery proportion rates. Various statistical methods controlling these error rates are described. The advantages and disadvantages of these methods are discussed. These methods are applied to gene expression data from two microarray studies and the properties of these multiple testing procedures are compared.

Keywords

multiple testing; familywise error rate; false discovery rate; false discovery proportion; adjusted p -value; resampling; microarray

1 Introduction

During the past several years, researchers have increasingly used large-scale multiple testing procedures for microarray data analysis involving thousands of genes. Various statistical procedures to address multiple testing problems associated with microarray studies have been reported. Microarray assays generate mRNA transcript concentration estimates for thousands of genes simultaneously.^{1–4} Some reviews of microarray data analysis can be found in Nature Genetics Supplement,⁵ Speed,⁶ and Parmigiani et al.⁷ One common goal in conducting microarray experiments is to identify genes that are differentially expressed, i.e., genes whose expression values are associated with a response or covariate of interest. The response could be censored survival time or other clinical outcomes, the covariates could be either categorical (e.g. treatment/control status, cell type) or continuous (e.g. dose of a drug, time of growth). A full treatment of multiple testing problems for different types of microarray data is found in Dudoit et al.⁸ The present paper focuses on the comparison of

Address for correspondence: Yongchao Ge, Mount Sinai School of Medicine, Department of Neurology, One Gustave L. Levy Place, Box 1137, New York, NY, 10029, yongchao.ge@mssm.edu.

gene expressions between two groups, as seen in the two microarray datasets^{9,10} used for illustration.

The expressions for m genes and n samples from a microarray experiment are typically represented by an $m \times n$ matrix X . Here m is typically several thousands or tens of thousands, and n is usually less than one hundred due to the cost of microarray experiments and limitation on biological materials. We assume that the first n_1 samples and the remaining $n_2 = n - n_1$ samples are formed into two groups. For each gene, we can formulate a null hypothesis test that gene i is non-differentially expressed,

H_i : The mean expression values of gene i are the same between the two groups.

The null hypothesis H_i can be tested by a two-sample Welch t -statistic,¹¹

$$t_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{\sqrt{\frac{s_{i2}^2}{n_2} + \frac{s_{i1}^2}{n_1}}},$$

where $\bar{x}_{i1}$ and $\bar{x}_{i2}$ denote the sample average expression values of gene i in the two groups, respectively; s_{i2}^2 and s_{i1}^2 are the corresponding sample variances.

The greater t_i is in absolute value, the stronger evidence is that gene i is differentially expressed. The p -values p_i , $i = 1, \dots, m$, can be defined as,

$$p_i = P(|T_i| \geq |t_i| | H_i).$$

This probability can be computed by a resampling method (permutation or bootstrap) or by a Student t -distribution table with an appropriate number of degrees of freedom. When the Student t -distribution table is applied to microarray data of small sample size, one needs to be comfortable with the assumption that x_{ij} are normally distributed. In either case, the p -values obtained can be used to rank the genes and the top ranking genes can be examined. Whether a gene is truly differentially expressed can be confirmed by carrying out further experiments such as quantitative real time PCR. However, one may want a quantitative assessment of the evidence concerning the possibly differentially expression genes in order to avoid follow up genes having little prospect of being truly differentially expressed. To address this need, one can carry out a simultaneous test for the null hypothesis for each gene that the expression is non-differentially expressed. In any such testing situation, two types of errors can occur: a false positive, or type I error, is committed when a gene is falsely declared to be differentially expressed, and a false negative, or type II error, is committed when the test fails to identify a truly differentially expressed gene. The numbers of type I errors and type II errors are denoted by V and D respectively* in Table 1. Since a typical microarray experiment measures the expressions of thousands of genes simultaneously, one is faced with an extreme multiple testing problem. Special problems arising from the

*The number of type II errors was denoted by T by Benjamini and Hochberg.¹³ We replace it by D to remove the confusion with the test statistic T in this paper.

multiplicity aspect include defining an appropriate type I error rate, and devising powerful multiple testing procedures that control this error rate and that incorporate the *joint* distribution of the test statistics $T_1, \dots, T_m$.

This paper covers three commonly used criteria to define type I error rates: the familywise error rates,¹² the false discovery rates,¹³ and the false discovery proportion rates.¹⁴ Section 2 introduces the basic notions of multiple testing and gives the definitions of these three criteria. Sections 3 reviews simple procedures that do not require resamplings. Sections 4 presents resampling-based procedures. The multiple testing procedures of Sections 3 and 4 are applied to the gene expression datasets from two published microarray studies described in Section 5. And finally, Section 6 summarizes our findings and outlines open questions.

2 Multiple testing

2.1 Set-up

We will adopt the notations of Romano and Wolf¹⁵ to describe a multiple testing problem. We first assume that gene expression data matrix X belongs to a certain family of multivariate distributions $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$, and hypothesis H_i can be viewed as a subset ω_i of Ω . Here Ω can be a nonparametric model, a parametric model, or a semiparametric model. For the microarray dataset of this paper, we can define the set for hypothesis H_i ,

$$\omega_i = \{\theta: \mu_{i1}(P_\theta) = \mu_{i2}(P_\theta)\},$$

where $\mu_{i1}(P_\theta)$ and $\mu_{i2}(P_\theta)$ are the mean expression values of gene i for the two groups respectively. Let the full set of hypothesis indices be $M = \{1, \dots, m\}$. For any subset K contained by M , define the intersection null hypothesis $H_K = \cap_{i \in K} H_i$ to be

$$H_K: \theta \in \bigcap_{i \in K} \omega_i.$$

Define the set of truly non-differentially expressed genes to be

$$M_0 = M_0(\theta) = \{i: \omega_i \ni \theta\}.$$

The complement set of M_0 is denoted by $M_1 = M_1(\theta) = \{i: \omega_i \not\ni \theta\}$. Therefore, the full set M is decomposed by $M_0(\theta)$ and $M_1(\theta)$, the cardinality numbers of which are denoted by m_0 and m_1 , respectively, as summarized in Table 1. A multiple testing procedure, can be represented by a decision function

$$\delta = \delta(X) = (\delta_1(X), \dots, \delta_m(X)), \quad (1)$$

where $\delta_i(X)=1$ indicates that gene i is claimed to be differentially expressed (rejection), and $\delta_i(X)$ is equal to zero otherwise. For example, a common cutoff value c for all genes' t -

statistics has a decision function $\delta(X)(c) = I(|T_1| > c), I(|T_2| > c), \dots, I(|T_m| > c)$, where $I(\cdot)$ is the indicator function. The cut-off value c can be a gene specific, and/or data adaptive random variable. One may also use the p -values $(P_1, \dots, P_m)$ to construct a decision function $\delta(\cdot)$. For any given multiple testing procedure, the decision function $\delta(X)$ has a multivariate probability distribution associated with θ . In the rest of this paper, many functions and variables depend on the observed microarray data X and on the unknown parameter θ , both of which are omitted for the sake of simplicity unless we want to emphasize this dependency. For example, $\delta(X)$ will be abbreviated by δ , and $M_0(\theta)$ by M_0 . For a given δ , the total number of rejected hypotheses is

$$R = \sum_{i=1, \dots, m} \delta_i.$$

The number of type I errors (false positives or falsely claimed differentially expressed genes) is

$$V = \sum_{i \in M_0(\theta)} \delta_i, \quad (2)$$

and the number of type II errors is

$$D = \sum_{i \in M_1(\theta)} (1 - \delta_i).$$

U , S and W in Table 1 can be similarly defined. We introduce a random variable Q , the false discovery proportion^{13,14}

$$Q = FDP = \frac{V}{R} \cdot I(R > 0).$$

2.2 Different criteria to control type I errors

V and D are two types of errors we are concerned about. Only a few papers have addressed the problem of type II errors, for example, Genovese and Wasserman introduce the false non-discovery rates¹⁶ and Taylor et al. propose the ‘miss’ rate.¹⁷ The former needs the assumption that the test statistics $T_1, \dots, T_m$ are samples of a mixture model of two distributions, and the latter gives a conservative estimate of the ‘miss’ rate. At the time of this writing, we are not aware of a general solution to control type II errors for dependent data. This paper focuses on controlling type I errors, as seen in most current literature and as traditionally done in single hypothesis testing. In multiple testing problems, various ways of defining type I error rates are in the following:

- *Per-comparison error rate (PCER)*, the ratio of the expected number of type I errors to the number of hypotheses, i.e.,

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

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

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

- *Familywise error rate* (FWER), the probability of committing at least k type I errors,^{14,18} i.e.,

$$\text{FWER}_k = P(V \geq k).$$

The values of interesting k tend to be much less than m . Especially, FWER_1 is exactly the same as the classically defined familywise error rate (FWER). The well known Bonferroni correction controls the FWER for almost all types of datasets.

- *False discovery rate* (FDR). The expectation of the false discovery proportion.¹³

$$\text{FDR} = E(Q) = E\left(\frac{V}{R} \mid R > 0\right) \cdot P(R > 0).$$

When $m_0 = m$, we have $\text{FDR} = \text{FWER}$.

- *False discovery proportion rate* (FDPR), the probability that the false discovery proportion is at least γ ,

$$\text{FDPR}_\gamma = P(Q \geq \gamma).$$

If γ is restricted on the interval $(0, 1/m)$, then FDPR_γ is equal to FWER_1 .

- *Proportion of false positives* (PFP), the ratio of the expected false positives to the expected positives,¹⁹ i.e.,

$$\text{PFP} = E(V)/E(R).$$

This is another way to characterize the ratio V/R .

- *Positive false discovery rate* (pFDR). If interest is in estimating an error rate when positive findings have occurred, then the pFDR proposed by Storey²⁰ is appropriate. It is defined as the conditional expectation of the proportion of type I errors among the rejected hypotheses, given that at least one hypothesis is rejected,

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

Storey²⁰ gives a Bayesian interpretation of this definition. In a microarray data study involving thousands of genes, the probability that $R > 0$ is almost equal to one, and so the FDR and the pFDR are nearly equivalent.

Table 2 summarizes the error rates for controlling the number of false positives: PFER and FWER_k ; and the error rates for controlling the false discovery proportion: FDR and FDPR_γ . The above four error rates give the expectations and the exceeding probabilities for V and Q . PFER is equivalent to PCER, which is the same as single hypothesis testing. The PFP and the pFDR are two different ways besides the FDR to describe the expectation of the false discovery proportion Q . The FWER_k provides an overall sense of the number of false positives for confirmation studies; and the FDPR_γ is useful to explore the data in finding interesting genes for further experiments.

2.3 Weak control, exact control and strong control of a multiple testing procedure

A multiple testing procedure, which can be specified by nested decision functions $\delta^\alpha(\cdot)$ as in equation (1), is designed to ensure that one of the error rates defined in Section 2.2 is no greater than α , for $0 < \alpha < 1$. For example, the Bonferroni procedure can be specified by the decision function

$$\delta_i^\alpha = I(p_i \leq \alpha/m).$$

Here p_i is the p -value of testing hypothesis H_i , and p_i can be viewed as a *statistic* of microarray data matrix X . It is important to note that the expectations and probabilities in the definitions of these type I error rates are determined by the specification of θ or the probability distribution P_θ . It is always helpful to have some sense of the type I error rates under the *complete null* hypothesis H_M ($m_0 = m$ or $M_0 = M$), i.e., no genes are differentially expressed. A multiple testing procedure under H_M is said to exhibit *weak control*. For the FDPR_γ , the weak control is

$$\text{FDPR}_\gamma(H_M) = \max_{\theta \in H_M} P_\theta(Q \geq \gamma) \leq \alpha.$$

The definition above is dependent upon H_M and the corresponding decision function δ of the multiple testing procedure. A weak control is a minimal requirement for a multiple testing procedure. However, under the complete null H_M , we have $V = R$, so

$$\text{pFDR}(H_M) = \text{PFP}(H_M) = 1.$$

Thus weak control of the pFDR and the PFP is impossible. Furthermore, under the complete null H_M , we have $Q = I(V > 0)$ implied by $V = R$, so

$$\text{FWER}(H_M) = \text{FDR}(H_M) = \text{FDPR}_\gamma(H_M) = \max_{\theta \in H_M} P_\theta(V \geq 1),$$

where γ can be any value taking over the interval (0,1).

Weak control is a useful first step to find out if there are any interesting genes at all in the microarray data, but is not satisfactory if we expect that some genes are differentially expressed, which happens to be the case for most microarray datasets. We should be concerned about the error control under the set M_0 of true null hypotheses. This is called *exact control*, i.e.,

$$\text{FDPR}_\gamma(H_{M_0}) = \max_{\theta \in H_{M_0}} P_\theta(Q \geq \gamma) \leq \alpha.$$

As M_0 is unknown beforehand and it is the goal of the microarray experiment to find out a possible value of M_0 , we need *strong control* of the error rate to safeguard against all possible situations, i.e.,

$$\text{FDPR}_\gamma(\Theta) = \max_{\theta \in \Theta} P_\theta(Q \geq \gamma) \leq \alpha.$$

2.4 Estimating and controlling type I error rates

In developing a multiple testing procedure, the first step is to estimate the type I error rate under some idealized situation. For example, for a fixed rejection region $[0, c]$ of p -values, let V_c and R_c be the numbers of false positives and rejections, respectively. A direct computation of the FDR is impossible without knowing the exact value of θ , but the false discovery rate FDR for this rejection region $[0, c]$ can be intuitively estimated by mc/R_c as shown in the following,

$$\text{FDR}_c = E\left(\frac{V_c}{R_c} \cdot I(R_c > 0)\right) \approx \frac{E(V_c)}{R_c} = \frac{m_0 \cdot c}{R_c} \leq \frac{mc}{R_c}$$

Especially, when c is the i -th smallest p -value $p_{(i)}$, we have the estimated FDR to be

$$\hat{\text{FDR}}_{p_{(i)}} = \frac{m \cdot p_{(i)}}{i} \quad (3)$$

This point estimate is relatively simple, but computing the mean and variance of it is more challenging especially in the situations that we do not want to impose strong assumptions on the data. The difficulty arises because the probability distribution of the FDR estimate is dependent upon R_c , which relies on the characteristics of the non-null hypotheses. For results about the expectation of the FDR estimate under the assumption that the null p -

values are independently distributed as $U[0,1]$, readers are referred to Theorem 1 in Storey, Taylor and Siegmund.²¹

A point estimate of the type I error rate can be used to *suggest* a multiple testing procedure by rejecting a hypothesis when its corresponding estimate is no greater than the level α . In order for this suggested testing procedure to be valid, attention is required for the proof of strong control. For example, the procedure suggested by equation (3) is equivalent to the BH procedure.^{13,22} Benjamini and Yekutieli²³ gave the proof that the BH procedure provides strong control for the FDR when the joint distribution of the p -values $(p_1, p_2, \dots, p_m)$ satisfy positive regression dependency for the subset M_0 (PRDS),²³ a much weaker assumption than that the null p -values are independently distributed. The meaning that statistics T satisfy PRDS is that for any increasing set D and for each i in M_0 , $P(T \in D | T_i = t)$ is nondecreasing in t . A set D is said to be increasing if for any $x \in D$ and $y_i \leq x_i, i = 1, \dots, m$, then $y \in D$. Here T, x and y are m -dimensional vectors, with the i -th component denoted by T_i, x_i and y_i , respectively.

When a new algorithm for a type I error rate comes out, readers should pay attention to the distinction between point estimation and strong control. For strong control, readers also need to know the assumptions required. For example, the popular software SAM²⁴ is a good procedure to *estimate* the FDR. The proofs of strong control on the FDR by the SAM are currently only given when p -values are independent or when some types of asymptotic arguments are invoked.²¹

3 Procedures controlling type I error rates

3.1 Assumptions

We will focus on the frequentist approach of handling multiple testing. There are several explorations of multiple testing with empirical Bayes ideas. Interested readers are referred to the papers by Efron and his collaborators^{25,26}. In this paper, we assume that the marginal distributions of the test statistic T_i or the p -value P_i on the true null hypothesis can be specified. For example, the null p -values can be assumed to be uniformly distributed on the interval $[0,1]$ for most applications. The null distribution of T_i can be estimated by resampling (permutation or bootstrap). This forms the weakest assumption in microarray data. The following lists four commonly used assumptions.

- *BAS assumption*: Each null p -value has a marginal distribution as $U[0,1]$. It allows possible joint dependence structure among the p -values.
- *SEP assumption*: In addition to the BAS assumption, it further assumes that the joint distribution of null p -values is independent of the joint distribution of non-null p -values, i.e.,

$$P(P_i \leq x_i, i \in M_0; P_j \leq y_j, j \in M_1) = P(P_i \leq x_i, i \in M_0) \cdot P(P_j \leq y_j, j \in M_1)$$

- *IND assumption*: In addition to the SEP assumption, it further assumes that the null p -values $(P_i, i \in M_0)$ are independently distributed.

- *MIX assumption*: A stronger assumption than IND, it further assumes that the non-null p -values are independently identically distributed as $G(\cdot)$. The function $G(\cdot)$ may be restricted to be no less than the distribution function for $U[0,1]$ and may further be assumed to be a convex function over the interval $[0,1]$.

The SEP assumption puts neither restrictions on the dependence structure of null p -values ($P_i, i \in M_0$), nor restrictions on the joint structure of non-null p -values ($P_i, i \in M_1$). This is a much weaker assumption than IND. We tend not to use the strongest assumption MIX, as rarely do we have the distribution of the non-null test statistics, which depends on how the parameters for non-null depart from the null. Furthermore, differentially expressed genes rarely behave similarly, which is in conflict with the homogeneity assumption on the non-null hypotheses imposed by the MIX assumption.

For the microarray study, the IND assumption means that all non-differentially expressed genes are acting independently of each other; the SEP assumption means that the non-differentially expressed genes do not interfere with differentially expressed genes. The three assumptions (BAS, SEP, and IND) can be relaxed by applying probability inequalities in their respective proof of strong control of type I error rates. For example, the BAS assumption for most procedures can be relaxed to require that

$$P(P_i \leq x) \leq x \text{ for } x \in [0, 1].$$

The SEP assumption for the LR-FDPi and Holm-FDR procedures in Table 3 can be relaxed to require that

$$P(P_i \leq x | P_j, j \in M_1) \leq x \text{ for } i \in M_0, x \in [0, 1].$$

The IND assumption for the BH procedure can be relaxed to require positively regression dependent (PRDS) as mentioned at the end of Section 2.4. These weaker assumptions such as PRDS are still difficult to verify for a specific microarray dataset and there is no clear consensus which assumption is appropriate. We can always use procedures that rely only on the BAS assumption, but this approach may lose power to detect interesting genes as seen in the datasets of Section 5.

3.2 Single-step, step-down, step-up

Bonferroni²⁷ and Holm²⁸ procedures are two classic procedures that control the FWER. The Bonferroni procedure rejects H_i when $p_i \leq \alpha/m$. It is a *single-step* procedure as the rejection of a particular hypothesis H_i is independent of what the p -values might be taken for other hypotheses. By contrast, Holm is a step-down procedure. After p -values are first ordered such that $p_{(1)} \leq \dots \leq p_{(m)}$, a general description of step-down procedures is in the following,

Step-down procedure—For non-decreasing critical values $0 \leq c_1 \leq \dots \leq c_m \leq 1$, let r be the greatest index satisfying $p_{(1)} \leq c_1, \dots, p_{(r)} \leq c_r$. Reject the hypotheses $H_{(1)}, \dots, H_{(r)}$ if r exists; otherwise, i.e., $p_{(1)} > c_1$, reject no hypothesis.

A step-down procedure begins with the smallest p -value $p_{(1)}$ (most significant) and continues rejecting hypotheses as long as the p -value is no greater than its corresponding critical value. Holm is a step-down procedure with the critical values

$$c_i^{Holm} = \alpha / (m - i + 1).$$

The classical procedure that controls the FDR is BH,¹³ which is a step-up procedure and gives strong control under the PRDS condition,²³ a weaker assumption than IND. A step-up procedure is described as in the following,

Step-up procedure—For non-decreasing critical values $0 \leq c_1 \leq \dots \leq c_m \leq 1$, let r be the smallest index satisfying $p_{(m)} > c_m, \dots, p_{(r+1)} > c_{r+1}$. Define $r = m$ if $p_{(m)} \leq c_m$. Reject the hypotheses $H_{(1)}, \dots, H_{(r)}$ if $r \geq 1$; otherwise, i.e., $p_{(i)} > c_i$ for all i , reject no hypothesis.

A step-up procedure begins with the greatest p -value $p_{(m)}$ (least significant) and continues accepting hypotheses as long as the p -value is greater than its corresponding critical value. BH is a step-up procedure with

$$c_i^{BH} = \alpha \cdot i / m.$$

In Table 3, we list the critical values of some simple procedures that provide strong control of type I error rates under various assumptions.

4 Resampling-based methods

For microarray datasets involving thousands of genes, one may apply a procedure that requires a strong assumption (a modification of assumptions IND or SEP), which is difficult to verify for a specific microarray data. If one applies procedures with the weakest assumption BAS, then there may not be enough power to detect differentially expressed genes as shown in Section 5.

A more fruitful approach to handle the multiple testing problem is to use data adaptive cutoff values from the actual microarray data by resampling the test statistics. It is beyond this paper's scope to review resampling algorithms here. Readers are referred to the papers^{8,33–35} for more details. We will introduce two simplified resampling algorithms in this paper. One is the k -maxT procedure to control the FWER_k . The k -maxT is a single-step procedure, and is a modification of maxT.¹² This procedure is the same as Algorithm A* proposed by Korn et al.¹⁴ The other is the maxZ procedure^{36,37} to control the false discovery proportion for independent data. We will use resamplings to generalize the maxZ procedure for dependent data in controlling the FDPR_γ .

4.1 Permutation algorithm for the k -maxT procedure to control the FWER_k

For resampling procedures to control the FWER_k , it is usually more convenient to compute the *adjusted p -values*,¹² since it is computationally intensive to compute the cutoff value for

each specified α . After we compute the adjusted p -values, we reject any hypothesis whose adjusted p -value is no greater than α . The adjusted p -value gives us the flexibility to explore the data without having to recompute the cutoff value c_i all over again. The idea of the k -maxT procedure is to estimate the FWER_k under the complete null[†] H_M for any interesting rejection region $[-\infty, -|t_i|] \cup [|t_i|, +\infty]$ of t -statistics, i.e., k -maxT adjusted p -value is defined as

$$\tilde{p}_i = P(k - \max_{1 \leq j \leq m} |T_j| \geq |t_i| | H_M),$$

where $k - \max_{1 \leq j \leq m} |T_j|$ denotes the k -th greatest statistic of $|T_1|, \dots, |T_m|$. In order for these adjusted p -values to provide a strong control of the FWER_k , we require that H_M gives the least favorable situation configuration in computing the adjusted p -value, i.e., for any $x \geq 0$,

$$P(k - \max_{1 \leq j \leq m} |T_j| \geq x | H_M) \geq P_\theta(k - \max_{j \in M_0(\theta)} |T_j| \geq x) \text{ for any } \theta \in \Theta. \quad (4)$$

This condition is expected to hold in many situations as the left hand side is based on computing the k -th greatest variable from m variables, and the right hand side is based on computing the k -th greatest variable from a subset of the aforementioned m variables. One should realize that the inequality is not true in general as the two sides are based on two different probability distribution generating mechanisms. The condition in equation (4) is weaker than the monotonicity assumption required by Romano and Wolf.¹⁵ Their assumption can be satisfied in many situations as shown in their Theorem 1. For the problem of finding differentially expressed genes in microarray data, the inequality holds as the computation of the test statistic for each gene is not involved with any other gene. Therefore the adjusted p -values computed under the complete null H_M can guarantee a strong control of the FWER_k . The details of the k -maxT procedure are described in Algorithm 1.

4.2 Permutation algorithm for the maxZ procedure to control the FDP_{γ}

The idea of the maxZ procedure is to simulate the distribution of $\max_{c \geq 0} Z_c$ under the complete null H_M , where Z_c is a normalized V_c ,

$$Z_c = \frac{V_c - \mu_c}{\sigma_c}. \quad (5)$$

Here $V_c = V_c(H_M) = \sum_{j \in M} (|T_j| - c)$ is the number of false positives for a cutoff value c of test statistics; μ_c and σ_c are the mean and standard deviation of $V_c(H_M)$ respectively. Let $\chi_{1-\alpha}$ be the $1 - \alpha$ quantile of the probability distribution of $\max_{c \geq 0} Z_c | H_M$. A *simultaneous upper prediction bound*³⁶ of the false positives V_c can be estimated by,

[†]We assume that parameter space for H_M consists of a single point for the sake of simplicity; otherwise, we take a parameter from H_M that gives the least favorable configuration in computing the probabilities.

$$\tilde{B}_c^{1-\alpha} = \mu_c + \sigma_c \cdot \chi_{1-\alpha},$$

in the sense that

$$P(V_c \leq \tilde{B}_c^{1-\alpha} \text{ for all } c \geq 0 | H_M) \rightarrow 1 - \alpha.$$

In order to generalize the simultaneous bound to any $\theta \in \Theta$, i.e.,

$$P_\theta(V_c \leq \tilde{B}_c^{1-\alpha} \text{ for all } c \geq 0) \geq 1 - \alpha,$$

we need the assumption that H_M gives the least favorable configurations in computing the $1 - \alpha$ quantile of $\max_{c \geq 0} Z_c$, i.e., for any $x \geq 0$,

$$P\left(\max_{c \geq 0} Z_c \leq x | H_M\right) \leq P_\theta\left(\max_{c \geq 0} Z_c \leq x\right) \text{ for any } \theta \in \Theta. \quad (6)$$

This condition is expected to hold in many situations as $Z_c = \frac{V_c(H_M) - \mu_c}{\sigma_c}$ on the left hand side is always no smaller than $Z_c = \frac{V_c(\theta) - \mu_c}{\sigma_c}$ on the right hand side (readers should note that μ_c and σ_c are the same constants on both sides, but V_c is depending on θ , see equation (2)). As in the condition required for the k -maxT procedure of equation (4), we need to be careful that the two sides of the above inequality are computed under two different probability distribution generating mechanisms.

Algorithm 1

Permutation algorithm for single-step k -maxT to control the FWER_k

For the original data matrix X , compute the test statistics $t_1, \dots, t_m$.

For the b th permutation, $b = 1, \dots, B$:

- 1 Permute the n columns of the data matrix X .
- 2 Compute test statistics $t_1^b, \dots, t_m^b$ for each hypothesis.
- 3 Next, compute $|t_{(k)}^b|$, the k -th greatest statistic of $|t_1^b|, \dots, |t_m^b|$.

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

$$\tilde{p}_i = \frac{\#\{b : |t_{(k)}^b| \leq |t_i|\}}{B} \quad \text{for } i = 1, \dots, m$$

Applying Meinshausen's technique,³⁸ an improved bound of V_c can be obtained by

$$\tilde{V}_c^{1-\alpha} = R_c - \max_{\tau \geq c} (R_\tau - \tilde{B}_\tau^{1-\alpha}),$$

which is smaller or equal to $\tilde{B}_c^{1-\alpha}$. Here R_c is the total number of rejections, equal to $\sum_{j=1}^m I(|T_j| \geq c)$. We can then obtain the simultaneous upper prediction bound of the false discovery proportion Q ,

$$\tilde{Q}_c^{1-\alpha} = \frac{\tilde{V}_c^{1-\alpha}}{\max(R_c, 1)},$$

satisfying

$$P_\theta(Q_c \leq \tilde{Q}_c^{1-\alpha} \text{ for all } c \geq 0) \geq 1 - \alpha, \text{ for any } \theta \in \Theta.$$

$\tilde{Q}_c^{1-\alpha}$ is also called false discovery proportion confidence envelope³⁹. We can then reject any hypothesis whose test statistic is no less than $\hat{c}$, where

$$\hat{c} = \min\{c \geq 0: \tilde{Q}_c^{1-\alpha} \leq \gamma\}.$$

The computational details of the maxZ procedure are described in Algorithm 2. This rejection procedure gives a strong control of the FDPR_γ under the assumption specified by equation (6).

Remarks

1. We use the mean and standard deviation to normalize V_c .^{36, 37} In fact, we can apply any nondecreasing function $F_c(\cdot)$ to normalize the variable V_c , i.e., Z_c is redefined to be $F_c(V_c)$ and $\tilde{B}_c^{1-\alpha}$ is redefined to be $F_c^{-1}(\chi_{1-\alpha})$. Meinshausen³⁸ takes F_c to be an empirical distribution of V_c , which is analogous to the quantile normalization used in microarray data analysis.⁴⁰ Different choices of function F_c may affect the power of the multiple testing procedure.
2. By noting that $m_0 = \max_{c \geq 0} V_c$, we can obtain a conservative $1 - \alpha$ upper confidence bound for m_0 given by $\max_{c \geq 0} \tilde{V}_c^{1-\alpha}$. And $m - \max_{c \geq 0} \tilde{V}_c^{1-\alpha}$ is a conservative $1 - \alpha$ lower confidence bound for m_1 .

5 Microarray applications

5.1 Data sets

Apo AI knock-out experiment—The Apo AI experiment⁹ was carried out as part of a study of lipid metabolism and atherosclerosis susceptibility in mice. The apolipoprotein AI

(Apo AI) is a gene known to play a pivotal role in HDL metabolism, and mice with the Apo AI gene knocked out have very low HDL cholesterol levels. The goal of the experiment was to identify genes with altered expression in the livers of these knock-out mice compared to inbred control mice. The treatment group consisted of eight mice with the Apo AI gene knocked out and the control group consisted of eight wild-type C57Bl/6 mice. For the 16 microarray slides, the target cDNA was from the liver mRNA of the 16 mice. The reference cDNA came from the pooled control mice liver mRNA. Among the 6 356 cDNA probes, about 200 genes were related to lipid metabolism. In the end, we obtained a 6 356 $\times$ 16 matrix with 8 columns from the controls and 8 columns from the treatments. Differentially expressed genes between the treatments and controls are identified by two sample Welch t -statistics.

Leukemia study—One goal of Golub et al.¹⁰ was to identify genes that are differentially expressed in patients with two types of leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing $m = 6\,817$ human genes. The learning set comprises $n = 38$ samples, 27 ALL cases and 11 AML cases. The data were preprocessed and filtered as described in Ge et al.³³ and the processed data (available at the Bioconductor package **multtest**) were summarized by a 3 051 $\times$ 38 matrix. Differentially expressed genes between ALL and AML patients were identified by two sample Welch t -statistics.

5.2 Results

Previous analysis of these two datasets, especially on the FWER and the FDR can be seen in Ge et al.³³ To keep the results brief, we only consider $k = 1$ for FWER $_k$ (or FWER) procedures and $\gamma = 0.05$ for FDP R_γ procedures. In total, we have 11 procedures for multiple testing: Holm, Šidák, RS-FWER, and k -maxT to control the FWER $_k$; BH, BY, and Holm-FDR to control the FDR; and LR-FDPi, RS-FDPi, RS-FDPii and maxZ to control the FDP R_γ . The p -values required for the simple procedures listed in Table 3 are obtained by

permutations, where the Apo AI knock-out experiment takes all $B = \binom{16}{8} = 12\,870$ permutations and the leukemia dataset takes 100 000 random permutations. The resampling algorithms listed in Section 4 take the exact same number of permutations as computing the p -values.

For the Apo AI dataset, the k -maxT and maxZ procedures reject 8 genes and 6 genes respectively at $\alpha = 0.05$, no other method listed in Section 3 is able to reject any gene. This shows that extreme power obtained by incorporating the dependency among the test statistics. Even though the methods in Section 3 are easy to implement, these critical values are unlikely to be improved upon.³⁰ The loss of power of these methods is due to that the critical values c_i are too general.

The results for the leukemia dataset are plotted in Panels A, B and C of Figure 1. Resampling-based procedures (k -maxT and maxZ) are more powerful than the simple methods of Table 3. For these simple methods, as expected, one that requires stronger assumptions tends to be able to reject more genes at the same level α . Two simple methods

stand out: LR-FDPi and BH. LR-FDPi is similar to maxZ in controlling FDP_{γ} . BH is comparable to the maxZ procedure that controls the median FDP (see panel D and more discussion in the next paragraph).

In general, we observe that FWER procedures reject fewer genes than FDR procedures, which is due to the strong criterion specified by the FWER. A direct comparison of FDR and FDP_{γ} is inappropriate due to different interpretations of α . From the maxZ procedure, we obtained the simultaneous upper prediction bound $Q^{1-\alpha}$ at different level of $1 - \alpha$ (see panel D of Figure 1). $Q^{0.5}$ gives the upper bound of the median FDP. Recall that the FDR is trying to control the expectation of the FDP. Therefore, we can use the median FDP to represent the performance of maxZ in comparison with the FDR procedures. It seems that maxZ is much more powerful than the BY procedure, and more or less the same as the BH procedure, which needs a stronger assumption (PRDS) than the maxZ procedure. This also shows the power of incorporating the dependency among the dataset.

For the Apo AI dataset, $Q^{0.5}$ from maxZ rejects 8 genes at $\gamma = 0.05$, the same as the ones the k -maxT procedure rejected at $\alpha = 0.05$. These 8 genes have been verified to be related to the knocked out gene. For the leukemia dataset, many more genes are expected to be differentially expressed. In fact, the maxZ procedure gives 661 genes as a 99% lower confidence bound for m_1 for the leukemia dataset (6 genes for the Apo AI dataset).

6 Discussion

For a large dataset such as microarray, we have seen the extreme power achieved by resampling methods as shown by k -maxT and maxZ, which exploit the correlation structure presented in the data. It seems that the usefulness of the simple procedures of Section 3 is very limited in a large dataset, even with the recent exciting theoretical progress in this area led by Romano and Shaikh.³⁰ These simple methods are useful when the test statistics satisfy IND or SEP assumptions, such as BH and RS-FDPi procedures, which are difficult to verify. Owen⁴¹ shows that the dependence among the hypothesis tests greatly affect the variance of number of false positives, and may affect the validity of the multiple testing procedures that require the IND assumption.

The FDR has its advantages in finding interesting genes than the more stringent criterion FWER. The generalized version FWER_k faces another difficulty of picking up an appropriate k in exploring the data. We have implemented a resampling algorithm for each error rates listed in Table 2, other than the FDR. Existing resampling algorithm on the FDR such as Yekutieli and Benjamini (1999)⁴² still requires the SEP assumption. The difficulty of developing a resampling algorithm for the FDR is due to the unknown probabilistic distribution specified by the non-null hypothesis. The maxZ procedure gets around this problem by estimating the simultaneous upper prediction bound of V_c , which depends only on the true null hypotheses. Before a resampling algorithm for the FDR under general dependent data is developed, one practical solution is to use maxZ to control the median FDP, rather than the FDR (the expectation of the FDP).

The maxZ procedure gives investigators much more insight into the data than controlling the median FDP. It generates simultaneous bounds of the FDP at different level $1 - \alpha$ as shown

in panel D of Figure 1. This procedure also produces a confidence bound for m_0 . Another advantage of maxZ is that it can be applied to most data as long as the complete null H_M gives the least favorable configuration to compute the distribution of $\max_{c \geq 0} Z_c$. This assumption is even unnecessary as long as one can find a distribution that gives a $1 - \alpha$ quantile $\chi_{1-\alpha}$ such that

$$\min_{\theta \in \Theta} P_{\theta} \left(\max_{c \geq 0} Z_c \leq \chi_{1-\alpha} \right) \geq 1 - \alpha$$

The closer the left hand side is to $1 - \alpha$, the more power is achieved to detect differentially expressed genes. The disadvantages of the maxZ procedure are that it is computationally expensive and that it may require special handling for a new dataset. Different strategies to generate the null distribution of $\max_{c \geq 0} Z_c$ or to estimate the quantile $\chi_{1-\alpha}$ may influence the validity and the power of analysis.

In designing multiple testing procedures, especially a resampling-based one, we need to balance the power, the running time, and the required assumption for providing strong control. k -maxT and maxZ achieve a good balance among them. The implementation details are relatively straightforward and the running time is only a couple of minutes at a regular desktop computer for both data sets in this paper. The assumptions required for both procedures are minimal. The step-down version of k -maxT is not proposed as a general solution as it requires a stronger assumption and as the power achieved from the step-down is negligible for a large data set such as microarray. Meinshausen's procedure³⁸ is not proposed here as it requires much more computation and it may suffer from the discreteness of p -values as seen in the minP procedure.³³

Some recent work on the local false discovery rate and sample size for multiple testing either requires Bayesian interpretation of the data or requires the MIX (or IND) assumption. Interested readers can check the papers^{43,44} for details. A general solution to computing the sample size in dealing with multiple testing problem for general dependent data seems very challenging, as is developing methods to control type II error rates.

Multiple testing's aim is to draw a cutoff value to claim which genes are differentially expressed. This is useful when not many genes are expected to be differentially expressed. It is probably not interesting to know exactly where this cutoff is drawn if many genes are expected to be differentially expressed. Other than focusing on how to decide the cutoff value of t -statistics or p -values, more attention should be paid to the design of a powerful statistic, or a good ranking of the genes, which is very useful for investigators to use microarray as a screening tool to focus on the top candidate genes for further investigations.

Acknowledgments

We thank Xiaochun Li and Chi-Hong Tseng for their valuable comments on the manuscript. We thank the editors for helpful comments that have led to an improved paper. This work was supported by NIH grants U19 AI 62623, RO1 DK46943 and contract HHSN 26600500021C.

References

1. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*. 1996; 14:457–60. [PubMed: 8944026]
2. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995; 270:467–70. [PubMed: 7569999]
3. Lockhart DJ, Dong HL, Byrne MC, Follettie MT, Gallo MV, Chee MS, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*. 1996; 14:1675–80.
4. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnology*. 1997; 15(13):1359–67.
5. Nature Genetics Editors, editor. The Chipping Forecast II. *Nature Genetics Supplement*. 2003; 32:461–552.
6. Speed, TP., editor. Statistical analysis of gene expression microarray data. Boca Raton: Chapman & Hall/CRC; 2003.
7. Parmigiani, G.; Garrett, ES.; Irizarry, RA.; Zeger, SL., editors. The Analysis of Gene Expression Data. New York: Springer; 2003.
8. Dudoit S, van der Laan MJ, Pollard KS. Multiple testing Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*. 2004; 3(1):Article 13.
9. Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM. Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research*. 2000; 10(12):2022–29. [PubMed: 11116096]
10. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999; 286:531–7. [PubMed: 10521349]
11. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika*. 1938; 29:350–62.
12. Westfall, PH.; Young, SS. Resampling-based multiple testing: Examples and methods for p -value adjustment. John Wiley & Sons; 1993.
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*. 1995; 57:289–300.
14. Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: Application to high dimensional genomic data. *Journal of Statistical Planning and Inference*. 2004; 124:379–98.
15. Romano JP, Wolf M. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*. 2005; 100(469):94–108.
16. Genovese C, Wasserman L. Operating characteristics and extensions of the false discovery rate procedure. *J R Statist Soc B* 2002. 64:499–517.
17. Taylor J, Tibshirani R, Efron B. The ‘miss rate’ for the analysis of gene expression data. *Biostatistics*. 2005; 6(1):111–7. [PubMed: 15618531]
18. Lehmann EL, Romano JP. Generalization of the familywise error rate. *The Annals of Statistics*. 2005; 33:1138–54.
19. Fernando RL, Nettleton D, Southey BR, Dekkers JCM, Rothschild MF, Soller M. Controlling the proportion of false positives in multiple dependent tests. *Genetics*. 2004; 166:611–9. [PubMed: 15020448]
20. Storey JD. A direct approach to false discovery rates. *J R Stat Soc B*. 2002; 64:479–98.
21. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc B*. 2004; 66:187–205.
22. Seeger P. A note on a method for the analysis of significance en masse. *Technometrics*. 1968; 10(3):586–93.

23. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple hypothesis testing under dependency. *The Annals of Statistics*. 2001; 29(4):1165–88.
24. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci*. 2001; 98:5116–21. [PubMed: 11309499]
25. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*. 2001; 96(456):1151–1160.
26. Efron, B. Microarrays, Empirical Bayes, and the two-groups model. Department of Statistics, Stanford University; 2006. <http://www-stat.stanford.edu/~brad/papers>
27. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilit. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936; 8:3–62.
28. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Statist*. 1979(6):65–70.
29. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*. 1967; 62:626–33.
30. Romano JP, Shaikh AM. Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics*. 2006a; 34(4):1850–73.
31. Romano, JP.; Shaikh, AM. On stepdown control of the false discovery proportion. *IMS Lecture Notes—Monograph Series: 2nd Lehmann Symposium—Optimality*; 2006b. p. 33-50.
32. Ge Y, Sealfon SC, Tseng CH, Speed TP. A Holm-type procedure controlling the false discovery rate. *Statistics and Probability Letters* 2007. 77:1756–62.
33. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. *Test*. 2003; 12(1):1–44. With discussion 44–77.
34. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Statistical Science*. 2003; 8(1):71–103.
35. van der Laan MJ, Dudoit S, Pollard KS. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Stat Appl Genet Mol Biol*. 2004; 3(1):Article 14.
36. Ge, Y.; Li, X.; Sealfon, SC.; Speed, TP. An upper prediction bound for the false discovery proportion. *Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]*; 2005; p. 2093-100.
37. Meinshausen N, Rice J. Estimating the Proportion of False Null Hypotheses among a Large Number of Independently Tested Hypotheses. *The Annals of Statistics*. 2006; 34(1):373–93.
38. Meinshausen N. False Discovery Control for Multiple Tests of Association under General Dependence. *Scandinavian Journal of Statistics*. 2006; 33(2):227–37.
39. Genovese C, Wasserman L. A stochastic process approach to false discovery control. *The Annals of Statistics*. 2004; 32:1035–61.
40. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003; 19:185–93. [PubMed: 12538238]
41. Owen AB. Variance of the Number of False Discoveries. *Journal of the Royal Statistical Society, Series B*. 2005; 67(3):411–26.
42. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*. 1999; 82:171–96.
43. Efron B. Large-scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*. 2004; 99 (465):96–104.
44. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate and sample size for microarray studies. *Bioinformatics*. 2005; 21:3017–24. [PubMed: 15840707]

Appendix: The definitions of D_1 , D_2 , D_3 in Table 3

There are no closed forms for D_1 , D_2 and D_3 . The computational details are in the following.

1. For $m_0 = k, k + 1, \dots, m$, define

$$S_1(k, m, m_0) = 1 + m_0 \sum_{1 \leq j \leq m_0 - k} \frac{k}{k+j} \cdot \left(\frac{1}{m_0 - j} - \frac{1}{m_0 - j + 1} \right),$$

and let

$$D_1 = D_1(k, m) = \max_{k \leq m_0 \leq m} S_1(k, m, m_0).$$

The above definition of D_1 is specific to the sequence $a_i = a_{i,k}$, where $a_{i,k}$ is defined in Table 3. For the general definition, readers are referred to the paper³¹.

2. We will use $g(j)$ to denote $\lfloor \gamma j \rfloor + 1$. Define a sequence $0 = a_1 \leq \dots \leq a_m = 1$, where a_i is $b_{\gamma i}$ as defined in Table 3. Let

$$S_2(\gamma, m, m_0) = m_0 a_1 + m_0 \sum_{\substack{0 \leq k < m-1 \\ m_0 \geq g(m-k)}} \frac{\alpha_{m-k} - \alpha_{m-k-1}}{\max(m_0 - k, g(m-k))}$$

and define

$$D_2 = D_2(\gamma, m) = \max_{1 \leq m_0 \leq m} S_2(\gamma, m, m_0).$$

3. Define $\beta_0 = 0$, $\beta_i = \frac{i}{\max\{m+i-\lceil \frac{i}{\gamma} \rceil + 1, m_0\}}$, $i=1, \dots, \lfloor \gamma m \rfloor$, and $\beta_{\lfloor \gamma m \rfloor + 1} = \frac{\lfloor \gamma m \rfloor + 1}{m_0}$. Let

$$N = \min\{\lfloor \gamma m \rfloor + 1, m_0, \left\lfloor \gamma \left(\frac{m-m_0}{1-\gamma} + 1 \right) \right\rfloor + 1\},$$

and set

$$S_3(\gamma, m, m_0) = m_0 \sum_{i=1}^N \frac{\beta_i - \beta_{i-1}}{i}.$$

Finally,

$$D_3 = D_3(\gamma, m) = \max_{1 \leq m_0 \leq m} S_3(\gamma, m, m_0).$$

Algorithm 2

Permutation algorithm for the maxZ procedure to control the FDP

Define a very dense grid of K critical values $0 = c_1 < \dots < c_K < \infty$. Compute the test statistics of the data matrix X , denote the ordered statistics by $|t_{(1)}| \leq \dots \leq |t_{(m)}|$.

- a** Permute the n columns of the data matrix X for B times. Compute test statistics $t_1^b, \dots, t_m^b$ for each hypothesis.
- b** For $k = 1, \dots, K$, compute the number of rejections with the cutoff value c_k ,

$$v_k^b = \sum_{i=1}^m I(|t_i^b| \geq c_k),$$

and the mean and standard deviation of $(v_k^1, \dots, v_k^B)$,

$$\mu_k = \frac{1}{B} \sum_{b=1}^B v_k^b, \quad \sigma_k = \sqrt{\frac{1}{B} \sum_{b=1}^B (v_k^b - \mu_k)^2}.$$

- c** The B samples of $\max_{c \in \mathcal{C}} Z_c |H_M$ are obtained by

$$z_{(1)}^b = \max_{k=1, \dots, K} \frac{v_k^b - \mu_k}{\sigma_k} \text{ for } b = 1, \dots, B.$$

- d** For $i = 1, \dots, m$, compute the mean $\mu_{(i)}^*$ and standard deviation $\sigma_{(i)}^*$ of $(v_{|t_{(i)}|}^b, b = 1, \dots, B)$, where $v_{|t_{(i)}|}^b = \sum_{j=1}^m I(|t_j^b| \geq |t_{(i)}|)$.
- e** Estimate the $1 - \alpha$ quantile $\chi_{1-\alpha}^1$ from the B samples $(z_{(1)}^1, \dots, z_{(1)}^B)$. The simultaneous upper confidence bound Q of the false discovery proportion for the rejection region $[-\infty, -|t_{(i)}|] \cup [|t_{(i)}|, +\infty]$ can be computed by,

$$\begin{aligned} \tilde{B}(i) &= \mu_{(i)}^* + \sigma_{(i)}^* \cdot \chi_{1-\alpha}^1 \\ \tilde{V}(i) &= i - \max \left(0, \max_{j=1, \dots, i} (j - \tilde{B}(i)) \right) \\ \tilde{Q}(i) &= \tilde{V}(i) / i \end{aligned}$$

- f** For a specific γ , find the maximum index k such that $\tilde{Q}_{(k)} \leq \gamma$ and reject the k hypotheses with the greatest t -statistics in absolute value. And use $\tilde{V}_{(m)}$ (or $m - \tilde{V}_{(m)}$) as an estimate of the $1 - \alpha$ upper confidence bound for m_0 (or $1 - \alpha$ lower confidence bound for m_1).

Figure 1.

Different multiple testing procedures for the Leukemia dataset. Each point of the curve in Panels A, B and C gives the minimum type I error rate (FWER_k , FDR or FDP_{γ}) achieved for the specific multiple testing procedure when the top 1, 2, ... genes are rejected. The curve for p -val gives the permutation based p -values. In Panel D, the maxZ procedure gives the simultaneous upper prediction bounds of the FDP at level $1 - \alpha = 0.50, 0.90, 0.95$ and 0.99 when different number of genes are rejected, while the BH and BY procedures are plotted in panel D to compare with the FDP bound at $1 - \alpha = 0.50$. All results are based on the same 100 000 random permutations.

Table 1

Summary table for multiple testing problems, based on Table 1 in Benjamini and Hochberg (1995).¹³

	accept	reject	total
true null hypotheses	U	V	m_0
false null hypotheses	D	S	m_1
total	W	R	m

Table 2Comparisons between the error rates for V and Q

random variable	expectation	exceeding probability
number of false positives V	$\text{PFER} = E(V)$	$\text{FWER}_k = P(V \geq k)$
false discovery proportion Q	$\text{FDR} = E(Q)$	$\text{FDPR}_\gamma = P(Q \geq \gamma)$

Table 3

Summary table for procedures controlling various type I error rates

error rate	procedure	critical value c_i^a	step direction	assumption
PCFR		α	single-step	BAS
FWER	Bonferroni ^b	α/m	single-step	BAS
	Holm ^c	$\alpha/(m - i + 1)$	step-down	BAS
	Šidák ^d	$1 - (1 - \alpha)^{1/(m-i+1)}$	step-down	IND
FWER _k	LR-FWER ^e	$\alpha \cdot a_{i,k}$	step-down	BAS
	RS-FWER ^f	$\alpha \cdot a_{i,k}/D_1(k, m)$	step-up	BAS
FDR	BH ^g	$\alpha \cdot i/m$	step-up	IND
	BY ^h	$\alpha \cdot (i/m)/C_m$	step-up	BAS
	Holm-FDR ^{i,j}	$\min\left\{\frac{m\alpha}{(m-i+1)^2}, 1\right\}$	step-down	SEP
FDRP _{γ}	LR-FDPi ^e	$\alpha \cdot b_{i,\gamma}$	step-down	SEP
	LR-FDPii ^e	$\alpha \cdot b_{i,\gamma}/C_{\lfloor m\gamma \rfloor + 1}$	step-down	BAS
	RS-FDPi ⁱ	$\alpha \cdot b_{i,\gamma}/D_3(\gamma, m)$	step-up	BAS
	RS-FDPii ^f	$\alpha \cdot b_{i,\gamma}/D_2(\gamma, m)$	step-down	BAS

$$a_{i,k} = \begin{cases} k/m & i \leq k \\ k/(m-i+k) & i > k \end{cases} \quad b_{i,\gamma} = \frac{\lfloor i\gamma \rfloor + 1}{m - i + \lfloor i\gamma \rfloor + 1}$$

$\lfloor x \rfloor$ is the greatest integer $\leq x$ $\lceil x \rceil$ is the smallest integer $\geq x$

^a $C_i = \sum_{j=1}^i 1/j$ D_1, D_2, D_3 are defined in the Appendix

^b Bonferroni (1936)²⁷, improved by Holm procedure.^c Holm (1979)²⁸^d Šidák (1967)²⁹^e Lehmann and Romano (2005)¹⁸. LR-FDPi is improved by RS-FDPii; LR-FWER at $k = 1$ is equivalent to Holm.^f Romano and Shaikh (2006a)³⁰^g Benjamini and Hochberg (2005)¹³^h Benjamini and Yekutieli (2001)²³ⁱ Romano and Shaikh (2006b)³¹^j Ge et al. (2007)³²