Multiplicity Issues in Microarray Experiments

F. Bretz[1], J. Landgrebe[2], E. Brunner[3]
[1]B&SR, Novartis Pharma AG, Basel, Switzerland
[2]Abteilung Biochemie II, Universität Göttingen, Göttingen, Germany
[3]Abteilung Medizinische Statistik, Universität Göttingen, Göttingen, Germany

Summary

Objectives: Discussion of different error concepts relevant to microarray experiments. Review of some commonly used multiple testing procedures. Comparison of different approaches as applied to gene expression data.

Methods: This article focuses on familywise error rate *(FWER)* and false discovery rate *(FDR)* controlling procedures. Methods under investigation include: Bonferroni-type methods and their improvements (including resampling approaches), modified Bonferroni methods, data-driven approaches, as well as the linear step-up method and its modifications. Particular emphasis lies on the description of the assumptions, advantages and limitations for the investigated methods.

Results: *FWER* controlling procedures are often too conservative in high dimensional screening studies. A better balance between the raw *P*-values and the stringent *FWER*-adjusted *P*-values may be required in many situations, as provided by *FDR* controlling and related procedures.

Conclusions: The questions remain open, which error concept to apply and which multiple testing procedure to use. Although we believe that the *FDR* or one of its variants will be applied more often in the future, long-term experience with microarray technology is missing and thus the validity of appropriate multiple test procedures cannot yet be assessed for microarray data analysis.

Keywords

Familywise error rate, false discovery rate, error concepts, screening

Methods Inf Med 2005; 44: 431–7

1. Introduction

The design of standard microarray experiments has a large impact on any related statistical inference. Multiplicity issues play a particularly important role: Due to the large number of variables, adequate statistical inference tools, which control the number of false-positives, are needed. Assume, for example, that the expression of 30,000 genes is investigated in a simple two-sample layout, comparing wildtype with a mutant. If only a few genes are significantly differentially expressed under both conditions and if further for each gene an appropriate two-sided two-sample test is performed at the significance level of 5%, we expect to obtain roughly 1,500 false-positives. In practice, it is unknown, how many and which of the statistically significant genes are truly positive. These 1,500 genes would have to be investigated in follow-up studies, exceeding any reasonable time and budget constraints. Thus, error concepts and test procedures, which control the number of false-positives at an acceptable level, have to be applied.

This article is concerned with multiple testing procedures (MTPs), which test $m > 1$ hypotheses while controlling an appropriate error rate at a pre-specified level α. From a multiple testing point of view, microarray experiments are mainly characterized through (i) large values of m, which can easily be in the 10,000s, (ii) high-dimensional distributions with unknown correlations, and (iii) a large proportion of true null hypotheses. In the following we review some MTPs with a focus on (i)-(iii).

2. Error Concepts

Let m denote the number of (null) hypotheses $H_1, ..., H_m$ to be tested. Let $M = \{1, ..., m\}$ denote the associated index set and denote the set of m_0 true hypotheses by $M_0 \subseteq M$, $m_0 = |M_0|$. In any testing situation, three types of errors can be committed. False-positives (negatives) occur when a true (false) null hypothesis is rejected (retained). In the hypothesis testing environment, these errors are denoted as *type I* and *type II errors*, respectively. *Type III errors* (correct rejection of a null hypothesis with a wrong directional decision) are usually of minor importance in microarray experiments and are thus not considered further in this article.

The related notation is summarized in Table 1. The number of type I errors is denoted by V and the number of rejected hypotheses is denoted by R. Note that R is an observable random variable, $S, T, U,$ and V are all unobservable random variables, while m and m_0 are fixed numbers, where m_0 is unknown.

A standard approach in univariate hypothesis testing ($m = 1$) is to choose an appropriate test, which maintains the type I error rate at a pre-specified level α. In multiple hypothesis testing several generalizations of the type I error rate are possible. The *per-comparison error rate* $PCER = E(V)/m$ is the expected proportion of type I errors among the m decisions (i.e., each test is conducted at level α, what amounts to ignoring the multiplicity problem altogether). The *familywise error rate* $FWER = P(V > 0)$ is the probability of committing at least one error. Finally, the *false discovery rate* $FDR = E(V/R \mid R > 0)\, P(R > 0)$ is related to (but not the same as) the expected proportion of false-positives among all significant results. Other error concepts exist and will be re-

viewed briefly later on. Due to the widespread use of the *FWER* and the *FDR* in microarray experiments, we restrict our attention to these two major error concepts. Note that the choice of the error control has to be done prior to the data analysis. In general, $PCER \leq FDR \leq FWER$ for a given MTP, since $V/m \leq 1_{\{R>0\}} V/R \leq 1_{\{V>0\}}$. Thus, a MTP which controls the *FWER* also controls the *FDR* and the *PCER*, but not vice-versa. *FWER* controlling procedures are therefore more conservative than *FDR* controlling procedures, leading to a smaller number of rejected hypotheses.

Before reviewing different MTPs, we introduce some more terminology. For any of the error concepts above, the error control is denoted as *weak* if the type I error rate is controlled only under the complete null hypothesis $H = \bigcap_{i \in M_0} H_i$, $M_0 = M$. For example, in the case of controlling the *FWER* weakly, $P(V > 0 \mid H) < \alpha$. If, for a given MTP, the type I error rate is controlled under any partial configuration $\emptyset \neq I \subseteq M$ of the $m_0 = |I| \leq m$ true null hypotheses, the error control is denoted as *strong*. Thus, in the case of controlling the *FWER* strongly,

$$\max_{I \subseteq M} P(V > 0 \mid \bigcap_{i \in I} H_i) < \alpha.$$

In microarray experiments, where it is unlikely that no gene is differentially expressed, it seems particularly important to have a strong error control.

Similar to univariate hypothesis testing, it is desirable to compute adjusted *P*-values for a given MTP, which are directly compared with the pre-specified level α. An adjusted *P*-value $\tilde{p}_i$ is defined as the smallest significance level for which one still rejects H_i, given a particular MTP. In case of the *FWER*, $\tilde{p}_i = \inf\{\alpha \in (0, 1) \mid H_i \text{ is rejected at } FWER = \alpha\}$. The marginal (i.e., unadjusted) *P*-values p_i are denoted as *raw P*-values. Finally, a particular MTP is denoted as a *single-step* procedure if the rejection of a single hypothesis does not depend on the decision of any other hypothesis. Otherwise, the MTP is denoted as a *stepwise* procedure. Stepwise procedures are further distinguished into step-down and step-up procedures. Given a (fixed) sequence of hypotheses $H_{(1)} < \ldots < H_{(m)}$, step-down procedures start testing the hypothesis most likely to be rejected ($H_{(1)}$) and step down through the sequence while rejecting the hypotheses. The procedure stops at the first non-rejection (at $H_{(i)}$, say), and $H_{(1)}, \ldots, H_{(i-1)}$ are rejected. Step-up procedures start test ing $H_{(m)}$ and step up through the sequence while retaining the hypotheses. The procedure stops at the first rejection (at $H_{(i)}$, say), and $H_{(1)}, \ldots, H_{(i)}$ are rejected. For more details on the theory of multiple testing see, for example, Hochberg and Tamhane [1].

Table 1 Type I and type II errors in multiple hypothesis testing. m – total number of hypotheses; m_0 – number of true null hypotheses; R – number of rejected hypotheses; V – number of incorrectly rejected hypotheses; remaining variables explained in the text

Hypotheses	not rejected	rejected	
true	U	V	m_0
false	T	S	$m - m_0$
	W	R	m

3. *FWER* Controlling Procedures

3.1 Bonferroni-type Procedures

The standard single-step Bonferroni approach compares the raw *P*-values with α/m, or, equivalently, the hypothesis H_i is rejected if $\tilde{p}_i = \min(mp_i, 1) < \alpha$. The strong *FWER* control follows directly from Bonferroni's inequality:

$$P(V > 0) = P(\bigcup_{i \in m_0} \{\tilde{p}_i \leq \alpha\}) \leq$$

$$\sum_{i \in M_0} P(\tilde{p}_i \leq \alpha) \leq m_0 \alpha / m \leq \alpha$$

where the probability expressions are conditional on $\bigcap_{i \in M_0} H_i$. The Bonferroni approach is a simple yet conservative MTP and many improvements have been proposed. Holm [2], for example, proposed a step-down approach, which basically consists of repeatedly applying Bonferroni's inequality while testing the hypotheses in a data-dependent order. Let $p_{(1)} \leq \ldots \leq p_{(m)}$ denote the ordered unadjusted *P*-values with the associated hypotheses $H_{(1)}, \ldots, H_{(m)}$. Then, $H_{(i)}$ is rejected, if $p_{(j)} < \alpha/(m - j + 1)$, $j = 1, \ldots, i, ,$ i.e., if all hypotheses $H_{(j)}$ preceding $H_{(i)}$ are also rejected. Equivalently, the adjusted *P*-values for the Holm procedure are $\tilde{p}_{(i)} = \min\{1, \max[(m - i + 1)p_{(i)}, \tilde{p}_{(i-1)}]\}$. The Holm procedure is a stepwise approach and is by construction more powerful than the Bonferroni approach. In typical microarray experiments, however, where m_0/m is close to 1, there are no practical differences and both methods lead to virtually the same set of significant genes.

Further improvements of the Bonferroni approach are available. Shaffer [3] and others took logical constraints between the hypotheses into account. But since these procedures are very computer-intensive already for small values of m, they are not yet applicable in microarray experiments. A second improvement, which takes the stochastic dependencies between the *P*-values into account, has often been used in the microarray literature. For simplicity, we restrict the representation to single-step approaches. Extensions to stepwise approaches are described by Westfall and Young [4].

For the single-step approach consider the adjusted *P*-values $\tilde{p}_i = P(\min_{1 \leq j \leq m} P_j \leq p_i \mid H)$, which are based on the joint distribution of $\mathbf{P} = (P_1, \ldots, P_m)$. The related MTP rejects H_i if $\tilde{p}_i < \alpha$. If marginally $P_j \sim U[0, 1]$, the MTP controls the *FWER* exactly. However, a strong error control is only assured if the *subset pivotality* condition holds: $\mathbf{P}$ is said to have the subset pivotality property, if for all $I \subseteq M_0$ the joint distribution of $\{P_i, i \in I\}$ is identical under the restrictions H and $\{\bigcap_{i \in I} H_i\}$. The subset pivotality thus ensures that the distribution of any sub-vector of *P*-values does not depend on the truth or falseness of the hypotheses not considered by this sub-vector. This condition is sufficient to guarantee a strong *FWER* control, although is has to be verified from case to case. Many examples exist [4, 5], in which the violation of the subset pivotality condition leads to an inflated error level. The subset pivotality condition will typically hold if contrast test statistics are used for the comparison of several treatments. An example where the subset pivotality fails is testing whether the correlations of random variables are 0: In this case it can

be shown that the joint distribution of two large sample test statistics depends on the unknown value of a third correlation [4].

Usually, the joint distribution of **P** is unknown. The following resampling method is suggested to approximate the true distribution [4], where t_i^{obs} is the observed test statistic, $i = 1, ..., m$.

FOR $b = 1, ..., B$

(1) Permute the rows of the data matrix **X**, obtaining $\mathbf{X}^b$

(2) Based on $\mathbf{X}^b$, compute the test statistics $t_1^b, ..., t_m^b$

END

OUTPUT $\hat{\tilde{p}}_i = \sum_{b=1}^{B} I\left(|t_i^b| \geq |t_i^{obs}| \right) / B$

Note that the resampling step has to be done design-dependent in order to maintain the design structure defined through **X**, where we assume the genes to be arranged across the columns and the subjects across the rows of **X**. Note also that due to the subset pivotality condition it is sufficient to resample under H and that the truth or falsehood of the individual hypotheses need not to be known. In the multivariate case (as it is the case in microarray experiments), the entire observation vector is shuffled in order to maintain the correlation structure. Clearly, the resampling approach is more powerful than other Bonferroni-type approaches since the correlations within the data are used for inference.

3.2 Modified Bonferroni Procedures

Simes [6] proposed the following single-step modified Bonferroni procedure: Reject H, if there exists a $j = 1, ..., m$, such that $p_{(j)} < j\alpha/m$. Note that the Simes test does not yield assessments for the individual hypotheses and it is only possible to reject the complete null hypothesis H. The Simes test is more powerful than Bonferroni but it has the drawback that *FWER* control is proven only for particular correlation patterns, e.g., for independent test statistics or certain positive dependency structures [7]. Thus, the Simes procedure and its modifications have to be applied with care, since the correlations among the genes are typically unknown. We refer to Section 4 for a more detailed discussion, when *FDR* controlling procedures relying on Simes' inequality are introduced.

Hochberg [8] proposed a step-up extension of the Simes procedure. Let $\tilde{p}_{(i)} = \min\{1, \min[(m - i + 1)p_{(i)}, \tilde{p}_{(i+1)}]\}$ denote the adjusted P-values. The related MTP then rejects $H_{(i)}$ if $\tilde{p}_{(i)} > \alpha$. The Hochberg procedure can bee seen as a reversed Holm procedure, since it uses the same critical values as Holm but in a reversed testing order: $H_{(i)}$ is rejected, if there exists a $j = i, ..., m$, such that $p_{(j)} < \alpha/(m - j + 1)$. By construction, Hochberg is more powerful than Bonferroni, Holm and Simes. But again, in typical microarray experiments the power differences are marginal. As it is based on Simes' inequality, the Hochberg procedure is similarly restricted to certain correlation structures. An improved version by applying the closed test procedure on Simes' inequality has been derived by Hommel [9]. However, the Hommel procedure is typically not applied in microarray experiments due to the necessary intensive computations.

3.3 Data-driven Ordering Procedures without Multiplicity Adjustment

A different approach is to order the hypotheses in a fixed sequence $H_{(1)} < ... < H_{(m)}$ prior to the experiment (in contrast to the Holm and the Hochberg procedures, where the ordering is performed data-dependent). Now define $\tilde{p}_{(i)} = \max\{p_{(i)}, \tilde{p}_{(i-1)}\}$. It can be shown that the related MTP, which rejects $H_{(i)}$ if $p_{(i)} < \alpha$, controls the *FWER* strongly at level α [10]. Such an approach, however, requires the pre-specification of the hypotheses prior to the experiment, which is typically not feasible in microarray data analysis. To circumvent this problem, Kropf and Läuter [11] recently proposed a novel MTP, which is based on ordering the test statistics according to a suitably chosen (data-dependent) selector statistic, which is stochastically independent from the test statistic. This independence assumption ensures that one can then test each hypothesis at full level α according to the hypotheses order, where the non-rejection at any step renders further testing unnecessary. The approach is based on the theory of exact stable multivariate tests [12]. For simplicity we restrict the representation to the parametric one-sample problem. The original procedure [11] then orders the genes according to the decreasing weights $w_i = \sum_{j=1}^{n} x_{ij}^2$, $i = 1, ..., m$, where n is the number of replications in the single sample of interest. Starting with the gene associated with the largest weight, the procedure then steps through while performing one-sample t-tests, each at level α. As long as the procedure keeps rejecting, the associated hypotheses of no differential expression are rejected. The procedure stops as soon as $p_{(i)} > \alpha$. The advantage of the procedure is that no multiplicity adjustment is needed, while the *FWER* is strongly controlled.

The Kropf and Läuter procedure does not need the variance homogeneity assumption to maintain the size at level α. But the power depends markedly on the homogeneity of the variances. If the variances are unequal (as to be expected in typical microarray experiments), those non-differentially expressed genes with a high variance may lead to large values of the selector statistic. In such cases, the procedure stops too early, resulting in a loss of power. Several smoothening approaches have been investigated to increase the power [13]. A simple procedure has recently been proposed by Hommel and Kropf [14]: Their procedure steps through the ordered hypotheses by comparing $p_{(i)}$ with α/k for some pre-specified integer $k \geq 1$. The procedure then stops only after retaining k hypotheses. Thus, this improved procedure allows for some non-differentially expressed genes while paying for a small multiplicity adjustment. Comparisons of data-driven ordering procedures with competing methods can be found in [13, 14].

4. FDR Controlling Procedures

4.1 General Remarks

The *FDR* is defined as $FDR = E(Q)$, with $Q = V/R$ if $R > 0$ and $Q = 0$ otherwise [15].

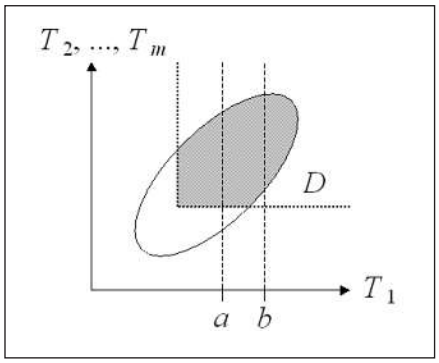

Fig. 1 Graphical illustration of the PRDS assumption. Explanations are given in the text.

Thus,

$$FDR = E(Q) = E(V/R \mid R>0)\, P(R>0) + 0 \cdot P(R=0) = E(V/R \mid R>0)\, P(R>0),$$

as stated in Section 2. Note that if $m_0 = 0$, then $V = 0$ and $FDR = 0$. If $m_0 = m$, then $FDR = E(1 \mid R>0)\, P(R>0) = P(R>0) = FWER$ and any *FDR* controlling MTP also controls the *FWER* weakly. Early ideas related to the *FDR* can be found in Seeger [16].

The *FDR* is a useful error concept, which suits particularly well to microarray analysis. *FWER* controlling procedures are typically too conservative for large number of hypotheses, i.e., the probability of missing differentially expressed genes is higher than for *FDR* controlling procedures. The later approaches address the error control in an intuitively more suitable way by considering the proportion of false-positives among all significant results. Caution has to be taken, however, when applying the *FDR*. First, as seen from the formula above, the *FDR* is not simply the expected proportion of false-positives among all significant results. This would be achieved by considering $E(V/R)$, which however is uncontrollable, since if $m_0 = m$, then $E(V/R) = 1$. Second, it is easily seen that the *FDR* can be reduced artificially by adding null hypotheses known to be false [17-19]. This is of particular importance in microarray experiments, since the inclusion of housekeeping genes or spiked-in genes is common practice. Third, it is worth pointing out that standard *FDR* controlling procedures do not provide information about the expected proportion of false-positives conditioned on having rejected at least one null hypothesis. We refer to Weller et al. [20] and the subsequent discussion in Zaykin et al. [21] for more details.

A number of alternative criteria have been introduced instead. The *positive FDR*, for example, is defined as $pFDR = E(V/R \mid R>0)$ [22, 23] and is closely related to the Empirical Bayes approach of Efron et al. [24]. A different concept is to control the proportion V/R directly: Korn et al. [25] and van der Laan et al. [26] independently introduced computer-intensive MTPs to control the *proportion of false positives* $PFP = P(V/R > \gamma)$, $0 < \gamma < 1$. We refer to the original articles for more details.

4.2 Linear Step-up Procedure

Benjamini and Hochberg [15] introduced the linear step-up (LSU) method described below, which in the meantime is widely used in microarray experiments. As before, let $p_{(1)} \leq \ldots \leq p_{(m)}$ denote the ordered *P*-values. If $k = \max\{i \mid p_{(i)} \leq i \cdot \alpha/m\}$ exists, reject $H_{(1)}, \ldots, H_{(k)}$. Equivalently, the adjusted *P*-values are given through $\tilde{p}_{(i)} = \min\{1, \min[mp_{(i)}/k, \tilde{p}_{(i+1)}]\}$. It follows from the proof given in Benjamini and Hochberg [15] that this MTP controls the *FDR* at level α or less, more specifically $FDR \leq m_0\alpha/m \leq \alpha$. In addition, the error control was only proven for independent test statistics. In the next paragraphs we discuss these issues in more detail and point to some extensions.

Several methods are available to estimate m_0 in order to apply the LSU method at level $\hat{m}_0\alpha/m$. Five methods, which estimate m_0, were compared in Hsueh et al. [27]. They concluded that the adaptive LSU method proposed by Benjamini and Hochberg [28] gives satisfactory empirical results. The latter considered the slopes of the lines passing the points $(m+1, 1)$ and $(i, p_{(i)})$ and take the lowest slope estimator to approximate m_0. The following adaptive procedure is thus proposed.

IF $p_{(i)} > i \cdot q/m$ for all i **THEN STOP**
ELSE COMPUTE

$$\hat{\beta}_i = \frac{1 - p_{(i)}}{m + 1 - i}, \quad i = 1, 2, \ldots$$

SET

$$j = \min\left\{i : \hat{\beta}_i < \hat{\beta}_{i-1}\right\}$$

$$\hat{m}_0 = \min\left\{\hat{\beta}_j^{-1} + 1,\ m\right\}$$

$$k = \max\left\{i \,\middle|\, p_{(i)} \leq i \cdot q/\hat{m}_0\right\}$$

REJECT $H_{(1)}, \ldots, H_{(k)}$

A second line of extending the LSU method focuses on the independence assumption mentioned above. Since the LSU method is closely related to Simes' test (see Section 3.2), similar concerns arise on the validity of the independence assumptions in practice. Benjamini and Yekutieli [29] showed that the LSU method controls *FDR* for certain positive dependency structures, to be specified now. A set D is called increasing, if $x \in D$, $y \geq x$, then $y \in D$. Benjamini and Yekutieli [29] then introduced the concept of a *positive regression dependency on a subset* (PRDS). An increasing set D is said to be PRDS on M_0, if for each $i \in M_0$, $P(\mathbf{X} \in D \mid X_i = x)$ is non-decreasing in x. The authors showed that the LSU method in fact controls *FDR* if the vector of test statistics $\mathbf{T} = (T_1, \ldots, T_m)$ is PRDS on M_0. The PRDS assumption holds in many practically relevant cases, in particular if $\mathbf{T}$ follows a multivariate normal distribution with non-negative correlations. But problems may already occur if all pairwise comparisons of three or more treatments are of interest. Figure 1 illustrates the PRDS assumption for the special case $M_0 = \{1\}$. In this example, $\mathbf{T}$ is positively correlated, with the first component being associated to the single true null hypothesis. If D denotes the positive orthant indicated by the dashed lines, it follows that $P(\mathbf{T} \in D \mid T_1 = a) < P(\mathbf{T} \in D \mid T_1 = b)$.

In cases where a negative correlation cannot be excluded prior to the analysis, Benjamini and Yekutieli [29] proposed a conservative modification of the LSU method using

$$\tilde{p}_{(i)} = \min\left\{1,\ \min\left[m\left(\sum_{j=1}^{m} \tfrac{1}{j}\right)p_{(i)}/k,\ \tilde{p}_{(i+1)}\right]\right\}.$$

This method is shown to control the *FDR* for any dependency structure (although it tran-

spires from Figure 2 that the conservativeness can be quite large). Other approaches exist, which take the correlations into account by relying on certain parametric assumptions. Yekutieli and Benjamini [30] proposed a resampling method to include the stochastic dependencies, Troendle [18] investigated asymptotic formulas and Somerville [31] provided exact critical values for step-down *FDR* procedures [32] and the LSU methods taking known correlations into account. Power comparisons between these and other competing methods can be found in Horn and Dunnett [33].

Table 2
Numerical comparison of different multiple testing procedures by computing the number of significant genes for different values of α using *multtest*. The abbreviations for the test procedures are explained in the text.

α	*rawp*	*Bonferroni*	*BH*	*BY*	*maxT*
0	0	0	0	0	0
0.1	2878	4	32	4	3
0.2	5398	6	60	4	5
0.3	7768	7	184	4	5
0.4	10290	9	334	6	8
0.5	12121	10	697	9	10

5. Application to Experimental Data

As an example, we show data generated in our laboratory. We compared the expression profile of a peripheral nerve in mice lacking a lysosomal membrane protein to wild-type mice (two-sample problem). We used 22k cDNA microarrays (i.e., each containing 22,000 genes) and material from six wild-type and six mutant animals. We performed twelve hybridizations (on 12 arrays) to compare the two animal groups by a common reference design, i.e., the targets were always labeled with Cy3 and the reference RNA (mouse liver) was labeled with Cy5. We performed two-sample *t*-tests to compare both conditions assuming homogeneous variances for each gene. In the following we compare several MTPs as applied to this data set. The methods under investigation are Bonferroni, Westfall and Young (abbreviated *maxT* hereafter), Benjamini and Yekutieli (*BY*), the LSU method (*BH*), as well as the unadjusted *P*-values (*rawp*). All calculations were performed using the R package *multtest* available from the Bioconductor website http://www.bioconductor.org. The Appendix includes the relevant code used for the following calculations.

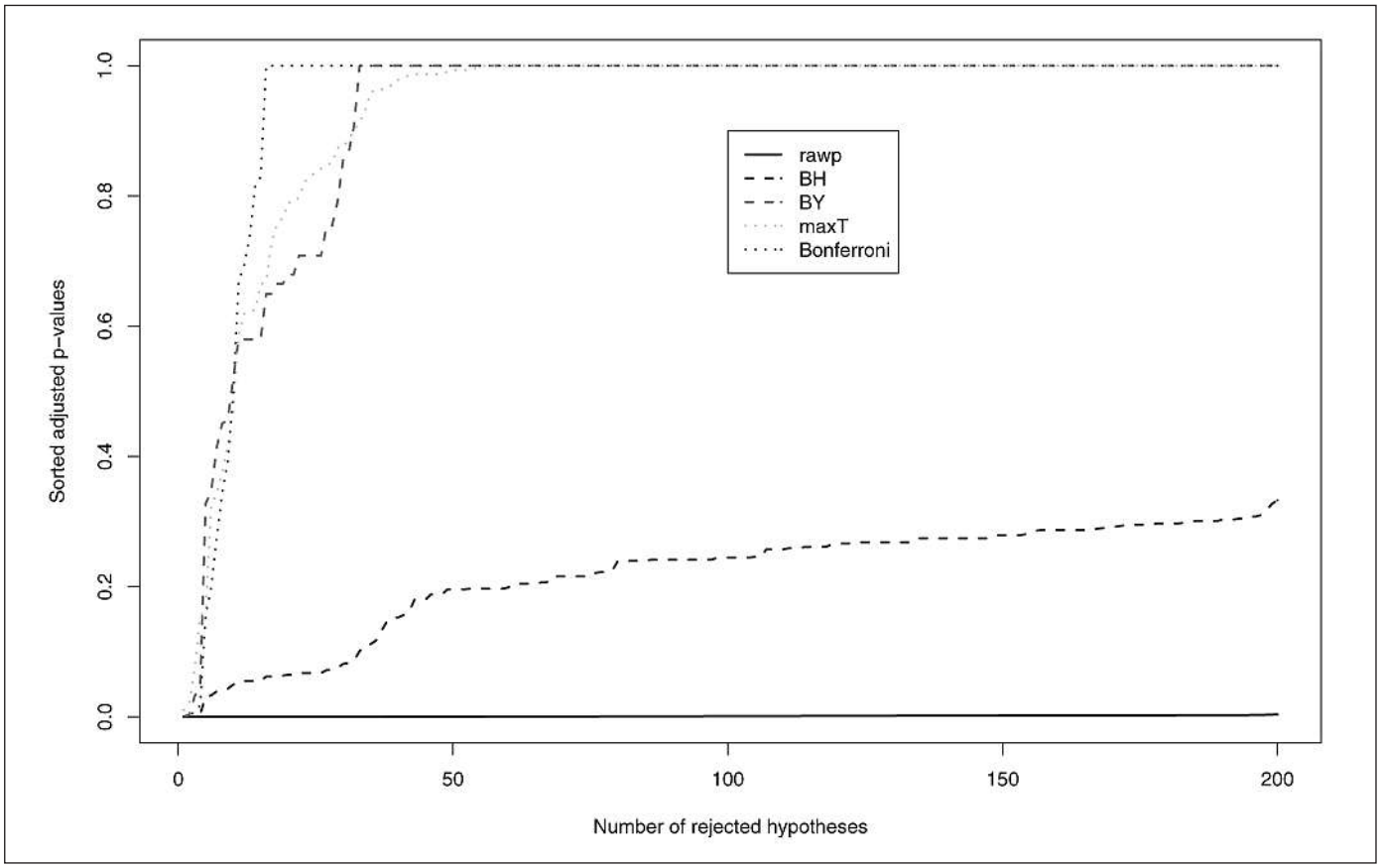

Fig. 2 Graphical comparison of different multiple testing procedures by plotting the sorted adjusted *P*-values for the 200 most significant genes in the experimental data using multtest. The abbreviations for the test procedures are explained in the text.

Table 2 shows the number of significant genes for different values of α. It transpires that simply considering *rawp* is inappropriate. For a significance level of 10%, for example, almost 3,000 genes are already statistically significant. The *BH* procedure greatly reduces the number of significant genes, while it is the most powerful method among the investigated MTPs. The remaining methods behave similar to each other. These findings are consistent with existing results from simulation studies published elsewhere [31, 33]. Note that the *BY* procedure tends to be overly conservative and should only be used if the PRDS assumption for the *BH* procedure is very questionable (e.g., when negative correlations or non-normal data are present). For the present example, *maxT* leads to a lower number of significances because of the small sample sizes (six replications per group). For this example, *multtest* considered the entire 12!/(6!6!) = 924 permutations, thus leading to a conservative approach, since due to the relatively small number of replications the nominal size is not fully exploited.

Figure 2 shows similar results in graphical form. For each MTP the number of rejected hypotheses are plotted against the sorted adjusted *P*-values. To simplify the graph, only the 200 most significant genes were plotted. The results are qualitatively similar to the previous findings.

6. Conclusions

It was our aim to focus the readers' attention to the fact that multiplicity adjustment plays a key role when analyzing microarray experiments. Not taking multiplicity issues into account may lead to a greatly inflated number of significant results, most of which are in fact false-positives. Thus, in the interest of the experimenter himself, procedures are required, which account for these issues.

In this paper we briefly reviewed some error concepts and multiple test procedures relevant to microarray experiments. It was one of our main goals to show that all of these methods are based on specific assumptions or at least have some particular characteristics and that the experimenter should be aware of them before applying a particular method. Further concepts and procedures exists, some of which might come to play a prominent role in the future. In particular, we refer to the series of articles by Dudoit et al. [34] and van der Laan et al. [26, 35], which discuss single-step and stepwise methods for controlling the $gFWER = P(V > k)$ for pre-specified k, the *PFP* and permutation methods not relying on the subset pivotality. We have also not covered the methods by Golub et al. [36] and Tusher et al. [37], which do not quite fit into the framework of this paper. We refer to the discussion in Dudoit et al. [5] and Ge et al. [38] instead.

Finally, we leave the question open, which error concept and multiple testing procedure to apply. We believe that the *FDR* or one of its variants will be applied more often in the future, although long-term experience with microarray technology is missing. Future research will help to assess the validity of the appropriate error concepts and test procedures for microarray data analysis.

References

1. Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: Wiley; 1987.
2. Holm S. A simple sequentially multiple test procedure. Scand J Stat 1979; 6: 65-70.
3. Shaffer JP. Modified sequentially rejective multiple test procedures. J Am Stat Assoc 1986; 81: 826-31.
4. Westfall PH, Young SS. Resampling-based multiple testing. New York: Wiley; 1993.
5. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci 2003; 18: 71-103.
6. Sarkar SK, Chang CK. Simes' method for multiple hypothesis testing with positively dependent test statistics. J Am Stat Assoc 1997; 92: 1601-8.
7. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986; 73: 751-4.
8. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988; 75: 800-2.
9. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 1988; 75: 383-6.
10. Westfall PH, Krishen A. Optimally weighted, fixed sequence and gatekeeping multiple testing procedures. J Stat Plan Infer 2001; 99: 25-40.
11. Kropf S, Läuter J. Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. Biom J 2002; 44: 789-800.
12. Läuter J, Glimm E, Kropf S. Multivariate tests based on left-spherically distributed linear scores. Ann Stat 1998; 26: 1972-88.
13. Westfall PH, Kropf S, Finos L. Weighted FWE controlling methods in high-dimensional situations. In: Recent developments in multiple comparisons procedures. Benjamini Y, Bretz F, Sarkar S (eds.). IMS Lecture Notes – Monograph Series, Vol. 47, pp 143-54, Beachwood, Ohio, USA; 2004.
14. Hommel G, Kropf S. Tests for Differentiation in Gene Expression Using a Data-Driven Order or Weights for Hypotheses. Biometrical Journal 2005 (to appear).
15. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc B 1995; 57: 289-300.
16. Seeger P. A note on a method for the analysis of significance en masse. Technometrics 1968; 10: 586-93.
17. Finner H, Roters M. On the false discovery rate and expected type I errors. Biom J 2001; 43: 985-1005.
18. Troendle JF. Stepwise normal theory multiple test procedures controlling the false discovery rate. J Stat Plan Infer 2000; 84: 139-58.
19. Hsu JC, Chang JY, Wang T. Multiple comparisons in screening for differential gene expressions from microarray data. In: Screening. Dean A, Lewis S (eds.). New York: Springer-Verlag; 2004.
20. Weller JI, Song JZ, Heyen DW, Lewin HA, Ron M. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics 1998; 150: 1699-1706.
21. Zaykin DV, Young SS, Westfall PH. Using false discovery rate approach in the genetic dissection of complex traits: a response to Weller et al. Genetics 2000; 154: 1917-8.
22. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. J Roy Stat Soc B 2004; 66: 187-205.
23. Storey JD. A direct approach to false discovery rates. J Roy Stat Soc B 2002; 64: 479-98.
24. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 2001; 96: 1151-60.
25. Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: Application to high dimensional genomic data. J Stat Plan Infer 2004 (in press).
26. van der Laan MJ, Dudoit S, Pollard KS. Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives. Statistical Applications in Genetics and Molecular Biology 2004; 3: June 15.
27. Hsueh HM, Chen JJ, Kodell RL. Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. J Biopharm Stat 2003; 13: 675-89.

28. Benjamini Y, Hochberg Y. The adaptive control of the false discovery rate in multiple hypotheses testing with independent statistics. J Educ Behav Stat 2000; 25: 60-83.
29. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple hypothesis testing under dependency. Ann Stat 2001; 29: 1165-88.
30. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plan Infer 1999; 82: 171-96.
31. Somerville PN. FDR step-down and step-up procedures for the correlated case. In: Recent developments in multiple comparisons procedures. Benjamini Y, Bretz F, Sarkar S (eds.). IMS Lecture Notes – Monograph Series, Vol. 47, pp. 100-18; Beachwood, Ohio, USA; 2004.
32. Benjamini Y, Liu W. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. J Stat Plan Infer 1999; 82: 163-70.
33. Horn M, Dunnett CW. Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. In: Recent developments in multiple comparisons procedures. Benjamini Y, Bretz F, Sarkar S (eds.). IMS Lecture Notes – Monograph Series, Vol. 47, pp. 48-64; Beachwood, Ohio, USA; 2004.
34. 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 (June 14).
35. van der Laan MJ, Dudoit S, Pollard KS. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. Statistical Applications in Genetics and Molecular Biology 2004; 3 (June 9).
36. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 1999; 286: 531-7.
37. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to ionizing radiation response. Proc Ntl Acad Sci 2001; 98: 5116-21.
38. Ge Y, Dudoit S, Speed T. Resampling-based multiple testing for microarray data analysis. Test 2003; 12: 1-77.

Correspondence to:
Frank Bretz
WSJ-27.1.005
Novartis Pharma AG
Lichtstr. 35, 4002 Basel, Switzerland
E-mail: frank.bretz@novartis.com

Appendix

In the following we present the R code used for the calculation of the example in Section 5.

```
library(multtest)

dget("C:/temp/1x2MmCRfix.put") -> data
data.cl <- c(0,0,0,0,0,0,1,1,1,1,1,1)
data.gnames <- rownames(data)

teststat <-
  mt.teststat(data,data.cl,test="t.equalvar")
df <-
  12-apply(data,1,function(x) sum(is.na(x)))
stat <- cbind(teststat,df)
stat <- stat[!is.na(stat[,1]),]

o <- order(abs(stat[,1]), decreasing = TRUE)
stat <- stat[o,]
rawp <-
  2 * (1 - pt(abs(stat[,1]),stat[,2]))

resT <- mt.maxT(data, data.cl,B=0)
#We are doing 924 complete permutations,
#924 = 12!/(6!6!)
maxT <- resT$adjp
maxT <- maxT[!is.na(maxT)]

procs <- c("Bonferroni","BH", "BY")
res <- mt.rawp2adjp(rawp, procs)
adjp <- res$adjp[order(res$index), ]
allp <- cbind(adjp, maxT)
allp <- allp[1:200,]

mt.reject(allp, seq(0, 0.5, 0.1))$r

dimnames(allp)[[2]] <-
  c(dimnames(adjp)[[2]],   "maxT")
procs <- dimnames(allp)[[2]]
procs <- procs[c(1, 3, 4, 5, 2)]
cols <- c(1, 2, 3, 5, 6)
ltypes <- c(1, 2, 2, 3, 3)

mt.plot(allp[, procs], stat[1:1000,1], plot-
  type = "pvsr", proc = procs, leg =c(100,
  0.9), lty = ltypes, col = cols, lwd = 2)
```