Journal of the Royal Statistical Society
SERIES B Statistical Methodology

*J. R. Statist. Soc.* B (2014)
**76**, *Part* 1, *pp.* 297–318

# Selective inference on multiple families of hypotheses

Yoav Benjamini

*Tel Aviv University, Israel*

and Marina Bogomolov

*Technion—Israel Institute of Technology, Haifa, Israel*

[Received May 2012. Revised March 2013]

**Summary.** In many complex multiple-testing problems the hypotheses are divided into families. Given the data, families with evidence for true discoveries are selected, and hypotheses within them are tested. Neither controlling the error rate in each family separately nor controlling the error rate over all hypotheses together can assure some level of confidence about the filtration of errors within the selected families. We formulate this concern about selective inference in its generality, for a very wide class of error rates and for any selection criterion, and present an adjustment of the testing level inside the selected families that retains control of the expected average error over the selected families.

*Keywords*: False discovery rate; Familywise error rate; Hierarchical testing; Multiple testing; Selective inference

## 1. Introduction

In modern statistical challenges we are often presented with a set of families of hypotheses, where the set and families may be large. In functional magnetic resonance imaging interest lies with the locations (voxels) of activation while a subject is involved in a certain cognitive task. The brain is divided into regions (either anatomic or functional), and the hypotheses regarding the locations in each region define a family (see, for example, Pacifico *et al*. (2004) and Benjamini and Heller (2007)). Searching for differentially expressed genes, the genes are often divided into gene sets, which are defined by prior biological knowledge. Each gene set defines a family of hypotheses (see Subramanian *et al*. (2005) and Heller *et al*. (2009)). In these examples, the families are clusters of units of interest: voxels or genes. Another problem having similar structure can be identified in multifactor analysis of variance, where for each factor interest lies with the family of pairwise comparisons between the levels of that factor. In more complex situations the set of hypotheses can be divided into families in different ways: an example of such research is the voxelwise genomewide association study (see Stein *et al*. (2010)), where the relationship between 448293 single-nucleotide polymorphisms (SNPs) and volume change in 31622 voxels (a total of $448293 \times 31622$ hypotheses) is explored across 740 elderly subjects. We may view this problem as a family for each gene, or a family for each voxel. This example is considered in detail in Section 7.

*Address for correspondence*: Yoav Benjamini, Department of Statistics and Operations Research, Sackler School of Exact Sciences and Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel.
E-mail: ybenja@post.tau.ac.il

 1369–7412/14/76297

In many such cases investigators tend to select promising families first, on the basis of the data at hand, and then look for significant findings only within the selected families. To avoid making erroneous discoveries within the selected families, it is common to apply a multiple-testing procedure in each selected family separately. In analysis of variance it is common first to select the significant factors, and then to perform *post hoc* tests (pairwise comparisons) within each selected factor by using Tukey's procedure. This controls the probability of making at least one erroneous discovery when applied in a single family of pairwise comparisons that is not subject to preselection.

Unfortunately, this strategy does not assure any level of confidence about the filtration of errors within the selected families. Let us demonstrate this by a simpler example.

### 1.1. Example 1

Facing families of 10 hypotheses each, where some families are all null and others include two non-null hypotheses, the investigator selects only the families that include a $p$-value which is below the Bonferroni threshold of 0.05/10 and rejects all such hypotheses in the family. Let us further assume that the $p$-values are independent and that the cumulative distribution function of the $p$-values of the non-null hypotheses is the square root of the uniform cumulative distribution function. Clearly, if an all-null family was selected, the probability of making a type I error is 1; if a non-all-null family was selected the conditional probability of rejecting a true null hypothesis or more in the family is still high: 0.23 (for the derivation see Appendix A). For an intuitive explanation that the conditional probability that a $p$-value of a true null hypothesis, say $P_j$, is below a threshold $x$ given that the minimal $p$-value in the family, $P_{(1)}$, is below $a$, note that

$$\Pr(P_j \leqslant x | P_{(1)} \leqslant a) = \frac{x}{\Pr(P_{(1)} \leqslant a)} \geqslant x \tag{1}$$

for any $x \leqslant a$, where the inequality is strict if $\Pr(P_{(1)} \leqslant a) < 1$. It follows that under this selection scheme the $p$-value under the null in a selected family is stochastically smaller than uniform on the interval $[0, a]$.

We address this problem of selective inference for a wide class of error rates. Quantifying differently the inflation of false discoveries when facing multiplicity, these error rates include the *per-family error rate* $E(V)$, where $V$ is the number of type I errors, the *familywise error rate* FWER, $\Pr(V \geqslant 1) = E(\mathbf{I}_{\{V \geqslant 1\}})$, the *false discovery rate* $\mathrm{FDR} = E(\mathrm{FDP})$, where $\mathrm{FDP} = V/R$ is the proportion of false discoveries out of all the discoveries $R$, defined as 0 if no discoveries are made (Benjamini and Hochberg, 1995), the false discovery exceedance $\mathrm{FDX} = \Pr(\mathrm{FDP} > \gamma) = E(\mathbf{I}_{\{\mathrm{FDP} > \gamma\}})$ for some prespecified $\gamma$ (see van der Laan *et al.* (2004) and Genovese and Wasserman (2006)) and the generalized error rates $k$-FWER, i.e. $\Pr(V \geqslant k) = E(\mathbf{I}_{\{V \geqslant k\}})$ (see van der Laan *et al.* (2004) and Lehmann and Romano (2005)) and $k$-$\mathrm{FDR} = E(\mathrm{FDP}\, \mathbf{I}_{\{V \geqslant k\}})$ (Sarkar, 2007). Although we emphasize in this work FDR and FWER, it addresses all the above error rates and others that can be written as $E(\mathcal{C})$ for some (random) measure of the errors performed $\mathcal{C}$. It does not address the Bayesian FDR, $\mathrm{Fdr} = E(V)/E(R)$ (Efron and Tibshirani, 2002) and the positive FDR, $\mathrm{pFDR} = E(V/R | R > 0)$ (Storey, 2003) that cannot be written as $E(\mathcal{C})$. See Farcomeni (2008) for a good review of multiple-error criteria, the relationship between them and different multiple-testing procedures.

When interest lies only with the selected families, we may wish to assure some level of confidence for the discoveries within the selected families. A natural request could be the control of

the expected value of some measure of errors $\mathcal{C}$ in each selected family $i$, i.e. control of $E(\mathcal{C}_i|i$ is selected). We have shown via example 1, where $\mathcal{C}_i = \mathbf{I}_{\{V_i \geqslant 1\}}$, that in some cases the common strategy of applying an $E(\mathcal{C})$ controlling procedure in each selected family separately results in $E(\mathcal{C}_i|i \text{ is selected}) = 1$. In general, the goal of such conditional control for any combination of selection rule and testing procedure and for any configuration of true null hypotheses is difficult to achieve.

We present a more modest goal about the errors within the selected families: the control of the expected average value of $\mathcal{C}$ over the *selected* families, where the average is 0 if no family is selected. In example 1, where $\mathcal{C} = \mathbf{I}_{\{V \geqslant 1\}}$, the expected average value of $\mathcal{C}$ over the selected families is the expected proportion of families with at least one erroneous rejection out of all the selected families.

Formally, let $\mathbf{P}_i$ be the set of $p$-values in family $i$, $i = 1, \ldots, m$, where the families are prespecified. Let $\mathbf{P}$ be the ensemble of these sets: $\mathbf{P} = \{P_i\}_{i=1}^m$. Let $\mathcal{S}$ be a selection procedure using as input the $p$-values $\mathbf{P}$, identifying the indices of the selected families. Define $|\mathcal{S}(\mathbf{P})|$, the number of selected families. The error criterion that interests us is

$$E(\mathcal{C}_{\mathcal{S}}) = E\left[\sum_{i \in \mathcal{S}(\mathbf{P})} \mathcal{C}_i \Big/ \max\{|\mathcal{S}(\mathbf{P})|, 1\}\right]. \tag{2}$$

When $\mathcal{S}(\mathbf{P}) \equiv \{1, \ldots, m\}$, i.e. when practically there is no selection and all the families are tested,

$$E(\mathcal{C}_{\mathcal{S}}) = E\left(\sum_{i=1}^{m} \mathcal{C}_i \Big/ m\right), \tag{3}$$

and therefore applying an $E(\mathcal{C})$ controlling procedure in each (such selected) family at level $q$ guarantees $E(\mathcal{C}_{\mathcal{S}}) \leqslant q$. It is easy to see that this is true also when the selection of families is done without relying on the data that are used for testing. However, $E(\mathcal{C}_{\mathcal{S}})$ may deteriorate when the selection depends on the data at hand, as already shown in example 1. The following example demonstrates how the extent of selection affects the expected average measure of errors over the selected families.

### *1.2. Example 2*

A family of $n$ hypotheses with corresponding $p$-values is selected if the minimum $p$-value in it is less than 0.05. Each selected family is tested by using a Bonferroni procedure at level $\alpha = 0.05$. Further assume that we have $m$ such families, and all the null hypotheses are true (with uniformly distributed independent $p$-values). Obviously, the expected value of averaged $\mathbf{I}_{\{V \geqslant 1\}}$ is controlled when the average is taken over all families. The expected value of the averaged $\mathcal{C} = \mathbf{I}_{\{V \geqslant 1\}}$ taken only over the selected families, namely $E(\mathcal{C}_{\mathcal{S}})$ (the average FWER hereafter) is given in Table 1 for various values of $m$ and $n$ (for the derivation see Appendix B).

We can immediately observe from the last column of Table 1 that in this example the average FWER over the selected families can climb high and reach above 0.5, whereas with no selection the level should be 0.05. It is also clear that the average FWER over the selected families increases when the extent of selection (presented in the third column) becomes more extreme. Note that in this example the average FWER is equivalent to $E(\mathcal{C}_{\mathcal{S}})$, where $\mathcal{C} = \text{FDP}$. Similar results were observed for the average PFER ($E(\mathcal{C}_{\mathcal{S}})$ for $\mathcal{C} = V$) rather than the average FWER.

The main result of this paper is that, to assure the control of $E(\mathcal{C}_{\mathcal{S}})$, we should control for $E(\mathcal{C}_i)$ in each selected family $i$ at a more stringent level: the nominal level $q$ should be multiplied

**Table 1.** Effect of selection on multiple testing in example 2, testing $m$ families with $n$ hypotheses in each†

| $m$ | $n$ | $E\{\lvert\mathcal{S}(\mathbf{P})\rvert/m\}$ | $E(\mathcal{C}_{\mathcal{S}})$ |
|---|---|---|---|
| 20 | 100 | 0.99 | 0.049 |
| 100 | 20 | 0.64 | 0.076 |
| 100 | 10 | 0.40 | 0.122 |
| 100 | 2 | 0.1 | 0.506 |

†All hypotheses are null. The selected set of families is $\mathcal{S}(\mathbf{P})$, including all families the minimum $p$-value in which is less than 0.05. Each selected family is tested by using the Bonferroni procedure at level 0.05, assuring that $E(\mathcal{C}_i) = E(\mathbf{I}_{\{V_i \geq 1\}}) \leqslant 0.05$. $\mathcal{C}_{\mathcal{S}}$ is the average number of families where at least one type I error was made. It can be seen that, as the selection becomes more stringent, the selection bias is more severe.

by the proportion of the selected families among all the families. This result, under some limiting conditions, is the focus of theorem 1 in Section 3. A general result of the same nature, covering more complicated selection rules, such as multiple-comparisons procedures that make use of plug-in estimators, is given in theorem 2.

## 2. Selective inference

Control on the average over the selected families is a manifestation of selective inference ideas that were developed in Benjamini and Yekutieli (2005). They made a distinction between simultaneous and selective goals in inference on multiple parameters, in the context of confidence intervals for the selected parameters. Simultaneous inference is relevant when the control of the probability that at least one confidence interval does not cover its parameter is needed. As a result, the simultaneous control also holds for any selected subset. However, when confidence intervals are built only for one set of selected parameters, the goal need not be that strict, and Benjamini and Yekutieli (2005) suggested a more liberal property: the control of the expected proportion of parameters not covered by their confidence intervals among the selected parameters, where the proportion is 0 if no parameter is selected (the false coverage statement rate FCR). Setting the goal of selective inference for the testing of multiple families and adopting the error measure in equation (2) is thus analogous to the goal of FCR. Thus the current work can be viewed as generalizing selective inference in two ways:

(a) the selected units are families of hypotheses rather than individual parameters or hypotheses, and the inference is made within the selected families;
(b) assessing errors within a family can be made by using a variety of error measures.

We shall now address these two points in detail.

We first illustrate the difference between selective inference on families of hypotheses and selective inference on individual hypotheses. If we obtain several families of hypotheses, taking the union of the families of hypotheses, treating them as a simple megafamily of hypotheses and controlling FDR globally for the combined set of discoveries is the manifestation of selective inference on the combined set of individual hypotheses. Controlling the expected average FDP

over the selected families of hypotheses is the manifestation of selective inference on families of hypotheses. Control of one does not imply the control of the other.

Assume that 40 families of hypotheses are selected. There are 36 families with one rejection in each, and there are no false discoveries in these families. In each of the remaining four families there are 10 rejections, five out of which are false discoveries. Thus in 36 selected families FDP $= 0$, whereas in the remaining four families FDP $= 0.5$. The average FDP over the selected families is $4 \times 0.5/40 = 0.05$. The total number of discoveries is 76, 20 out of which are false discoveries. Therefore FDP for the combined set of discoveries is $20/76 = 0.26$. Thus control of the expected average FDP over the selected families does not imply control of FDR for the combined set of discoveries.

Nor does the control of FDR for the combined set of discoveries guarantee control of the expected average FDP over the selected families. Assume that there are 20 selected families, each with one erroneous and one correct rejection, whereas in each of the other 20 families there are 18 rejections, all of which are correct. The total number of discoveries is 400, 20 out of which are false discoveries. Therefore, FDP for the combined set of discoveries is $20/400 = 0.05$. However, the average FDP over the selected families is $20 \times 0.5/40 = 0.25$.

We shall now show how the choice of the error measure $\mathcal{C}$ within the selected families is reflected in $E(\mathcal{C}_{\mathcal{S}})$, for the two most common choices of $\mathcal{C}$. When $\mathcal{C} = \mathbf{I}_{\{V>0\}}$, this error measure is the expected proportion of families with at least one type I error out of all the selected families. In this case it is similar to the overall false discovery rate that was defined in Heller *et al.* (2009) for use in microarray analysis (see Section 5). When $\mathcal{C} = \mathrm{FDP}$, the error measure in equation (2) becomes less stringent: it is the expected average FDP over the selected families. The difference between the average FDP and the proportion of families with at least one type I error may be very large. If three families are selected, with false discovery proportions equal to 0.04, 0.05 and 0.06, the average FDP is 0.05, whereas the proportion of families with at least one type I error is 1. The choice between these two error rates should be guided by the application. If we can bear some false discoveries in the selected families as long as the average FDP over the selected families is small, the control of the expected average FDP may suffice. Alternatively, if we wish to avoid even one false discovery in a selected family, control of the expected proportion of families with at least one type I error would be appropriate.

In many applications, controlling the expected average measure of errors over the selected families is simply a more appropriate measure of error for the interpretation of the results than controlling an error rate globally for the combined set of discoveries, because it gives some confidence in the discoveries within the selected families. This important point is illustrated with an application in Section 7. Even in the problems where no selection takes place, Efron (2008) argued that one should obtain control of an error rate in each family separately, implying control on the average (over all the families). When the selection takes place, control on the average (over the selected families) becomes even more important. Finally, in some cases power may be gained by controlling an error rate on the average rather than globally for the combined set of discoveries, even though this is not the motivating reason for our emphasizing the control on the average over the selected families.

## 3. Selection-adjusted testing of families

When all the families are selected with probability 1, no adjustment to the testing levels should be done because the average over the selected families is the average over all. As the selection rule is more stringent and tends to select fewer families, the adjustment should be more

severe. We start with the adjustment for simple selection rules and only then turn to the general case.

*Definition 1* (simple selection rule). A selection rule is called simple if for each selected family $i$, when the $p$-values not belonging to family $i$ are fixed and the $p$-values inside family $i$ can change as long as family $i$ is selected, the number of selected families remains unchanged.

A similar notion was defined by Benjamini and Yekutieli (2005) in the context of data-based parameter selection. Many selection rules are indeed simple. Any rule where a family is selected only on the basis of its own $p$-values is a simple selection rule, e.g. the rule that was used in example 1. In Section 5 we show that, when the selection of the families is done by using hypotheses testing, the widely used step-up and step-down multiple-testing procedures provide simple selection rules, even though the decision whether a family is selected or not depends on the $p$-values belonging to other families as well.

The following simple selection-adjusted procedure (*procedure 1*) offers the selection adjustment for a simple selection rule of families.

*Step 1*: apply the selection rule $\mathcal{S}$ to the ensemble of sets $\mathbf{P}$, identifying the selected set of families $\mathcal{S}(\mathbf{P})$. Let $R$ be the number of selected families (i.e. $R = |\mathcal{S}(\mathbf{P})|$).
*Step 2*: apply the $E(\mathcal{C})$ controlling procedure in each selected family separately at level

$$Rq/m.$$

*Theorem 1.* For any error rate $E(\mathcal{C})$ such that $\mathcal{C}$ takes values in a countable set, suppose that we have a testing procedure that can control $E(\mathcal{C})$ at any desired level $\alpha$ under the dependence structure of the $p$-values within a family. If the $p$-values in each family are independent of the $p$-values in any other family then for any simple selection rule $\mathcal{S}(\mathbf{P})$ the selection-adjusted procedure guarantees $E(\mathcal{C}_{\mathcal{S}}) \leqslant q$.

*Remark 1.* For all practical purposes $\mathcal{C}$ is a count or a ratio of counts, so the condition on the values that $\mathcal{C}$ takes is satisfied.

*Remark 2.* The adjustment for selection effect depends on the extent of the selection. As the proportion of selected families becomes smaller, the adjustment becomes more stringent. When all the families are tested, i.e. there is practically no selection, the quantity $Rq/m$ is equal to $q$, yielding that there is no adjustment needed, as we showed in Section 1. The intuition behind this adjustment concurs with the results of example 2, where we saw that $E(\mathcal{C}_{\mathcal{S}})$ increases as the expected proportion of selected families becomes smaller.

### 3.1. Proof of theorem 1

The idea of the proof is similar to the proof of theorem 1 in Benjamini and Yekutieli (2005). For each error criterion $E(\mathcal{C})$, let $\mathcal{C}_+$ be the countable support of $\mathcal{C}$. Since the selection rule is simple, we can define the following event on the space of all the $p$-values not belonging to family $i$: if family $i$ is selected, $k$ families are selected including family $i$. Denote this event by $C_k^{(i)}$. For any simple selection rule

$$E(\mathcal{C}_{\mathcal{S}}) = \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \sum_{c \in \mathcal{C}_+} c \Pr\{\mathcal{C}_i = c, i \in S(\mathbf{P}), C_k^{(i)}\}. \tag{4}$$

Note that the simple selection-adjusted procedure does not reject any hypothesis in families which are not selected. Therefore $\mathcal{C}_i = 0$ for each family $i$ that is not selected. Hence, for this procedure we obtain

$$E(\mathcal{C}_S) = \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \sum_{c \in \mathcal{C}_+} c \Pr(\mathcal{C}_i = c, C_k^{(i)})$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \sum_{c \in \mathcal{C}_+} c \Pr(\mathcal{C}_i = c) \Pr(C_k^{(i)}) \tag{5}$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} E(\mathcal{C}_i) \Pr(C_k^{(i)}). \tag{6}$$

Equality (5) follows from the independence between $P_i$ and the set of $p$-values not belonging to family $i$, for each $i = 1, \ldots, m$. In expression (6), for each $k$ and $i$, $\mathcal{C}_i$ is the value of random variable $\mathcal{C}$ in family $i$, when a valid $E(\mathcal{C})$ controlling procedure is applied at level $kq/m$ in each selected family. Since there are no rejections in families that are not selected, $\mathcal{C}_i$ takes the value 0 there, so $E(\mathcal{C}_i) \leqslant kq/m$ for each $i = 1, \ldots, m$. Now, using this inequality and the fact that $\Sigma_{k=1}^{m} \Pr(C_k^{(i)}) = 1$ for each $i = 1, \ldots, m$, we obtain

$$\sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} E(\mathcal{C}_i) \Pr(C_k^{(i)}) \leqslant \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \frac{kq}{m} \Pr(C_k^{(i)}) = q. \tag{7}$$

Results (6) and (7) complete the proof.

*Remark 3.* It follows from the proof of theorem 1 that we could actually use different adjustments in each family. Given the weights $w_i$, $i = 1, \ldots, m$, which do not depend on the data at hand and satisfy $\Sigma_{i=1}^{m} w_i = m$, we could apply $E(\mathcal{C})$ controlling procedure at level $Rqw_i/m$ in each family $i \in \mathcal{S}(\mathbf{P})$. If there is *a priori* knowledge that a certain family contains more false null hypotheses than other families, we may associate this family with a higher weight $w_i$, and then the adjustment in this family will be less stringent than in other selected families, thereby offering a gain in power.

Theorem 1 gives the adjustment of the testing level in each selected family which is sufficient for the control of $E(\mathcal{C}_S)$. We shall now show that in some special cases this adjustment is necessary, adapting example 6 in Benjamini and Yekutieli (2005) for our needs.

### 3.2. Example 3

Assume that all families tested are of equal size $n$, all null hypotheses are true and all the $p$-values are jointly independent and uniformly distributed. Let us order the families by their minimal $p$-values. The simple selection rule is to choose the $k$ families with the smallest minimal $p$-values. Assume that each selected family is tested by using the Bonferroni procedure at level $q'$. In this case the average error rate over the selected families is

$$E(\mathcal{C}_S) = E\left( \sum_{i \in \mathcal{S}(\mathbf{P})} V_i \right) \Big/ k.$$

Using the fact that the families where at least one erroneous rejection is made are the families with the smallest minimal $p$-values, it is easy to see that, if $m/k$ is not much larger than 1 (say $k = 3m/4$), we obtain

$$E\left( \sum_{i \in \mathcal{S}(\mathbf{P})} V_i \right) \Big/ k \approx E\left( \sum_{i=1}^{m} V_i \right) \Big/ k = mq'/k,$$

and the adjustment $q' = kq/m$ is necessary for assuring that $E(\mathcal{C}_S) \leqslant q$.

*Remark 4.* Example 3 shows that in some cases where all the null hypotheses are true the adjustment of the testing level at step 2 of procedure 1 is necessary. However, when there are families which consist of only false null hypotheses, an improvement of procedure 1 can be made. Since $\mathcal{C} = 0$ in the families which consist of only false null hypotheses, the first sum in equation (3) can be taken over the indices of families containing one or more true null hypotheses. Therefore, it follows from the proof of theorem 1 that the same result holds if each selected family is tested at level $Rq/(m - m_{11})$, where $m_{11}$ is the number of families consisting of only false null hypotheses. One could think of an adaptive version of procedure 1, where the parameter $m_{11}$ is estimated by using the data and then plugged in the testing level at step 2 of the procedure. However, we believe that in most applications $m - m_{11} \approx m$; therefore this improvement would not be significant.

Not all the selection rules are simple: examples are adaptive multiple-testing procedures, as noted in Section 5. Still a very similar procedure, the selection-adjusted procedure (*procedure 2*), even if somewhat more elaborate, offers selection adjustment for any selection rule. This procedure reduces to procedure 1 when the selection rule is simple.

*Step 1*: apply the selection rule $\mathcal{S}$ to the ensemble of sets $\mathbf{P}$, identifying the selected set of families $\mathcal{S}(\mathbf{P})$.
*Step 2*: for each selected family $i$, $i \in \mathcal{S}(\mathbf{P})$, partition the ensemble of sets $\mathbf{P}$ into $P_i$, the set of the $p$-values belonging to family $i$ and taking the values $p_i$, and $\mathbf{P}^{(i)}$ (the ensemble of sets $\mathbf{P}$ without the set $P_i$) and find

$$R_{\min}(\mathbf{P}^{(i)}) := \min_{p_i}\{|\mathcal{S}(\mathbf{P}^{(i)}, P_i = p_i)| : i \in \mathcal{S}(\mathbf{P}^{(i)}, P_i = p_i)\}, \tag{8}$$

the minimal number of selected families when family $i$ is selected and the $p$-values for other families do not change.
*Step 3*: for each selected family $i$, apply the $E(\mathcal{C})$ controlling procedure at level

$$R_{\min}(\mathbf{P}^{(i)})q/m.$$

*Theorem 2.* For any error rate $E(\mathcal{C})$ such that $\mathcal{C}$ takes values in a countable set, suppose that we have a testing procedure that can control $E(\mathcal{C})$ at any desired level $\alpha$ under the dependence structure of the $p$-values within a family. If the $p$-values in each family are independent of the $p$-values in any other family then for any selection rule $\mathcal{S}(\mathbf{P})$ the selection-adjusted procedure guarantees $E(\mathcal{C}_{\mathcal{S}}) \leqslant q$.

The proof is given in Appendix C.

*Remark 5.* We could guarantee that in each selected family at least one rejection is made by adding this requirement explicitly into the selection rule: 'Select the largest number of families so that all selected families will have a rejection when applying the selective inference adjustment at level $q$'. Interestingly, when each family consists of one hypothesis only, and each hypothesis within the selected family is rejected at level $\alpha$ if its $p$-value is less than $\alpha$, the selection-adjusted procedure is equivalent to the procedure in Benjamini and Hochberg (1995) (the BH procedure) applied on the set of $p$-values.

## 4. Average control under dependence across the families

We now consider the case where the set of all the $p$-values is positive regression dependent on the subset (PRDS) of true null hypotheses, as defined below.

First recall that a set $D$ in $\mathbb{R}^n$ is increasing or decreasing if $x \in D$ and respectively $y \geqslant x$ or $y \leqslant x$ implies that $y \in D$.

*Definition 2* (Benjamini and Yekutieli, 2001). The vector $\mathbf{X}$ is PRDS on $I_0$ if, for any increasing set $D$ and for each $i \in I_0$, $\Pr(\mathbf{X} \in D | X_i = x)$ is non-decreasing in $x$.

In addition, we require that the selection rule be concordant, as defined in Benjamini and Yekutieli (2005).

*Definition 3* (Benjamini and Yekutieli, 2005). A selection rule is concordant if, for each $i = 1, \ldots, m$ and $k = 1, \ldots, m$, $\{\mathbf{P}^{(i)} : k \leqslant R_{\min}(\mathbf{P}^{(i)})\}$ is a decreasing set.

It is easy to see that many selection rules are concordant. Both selecting each family where its minimum $p$-value is less than $q$ and selecting $k$ families with the smallest minimal $p$-values are concordant selection rules. When the selection is made via hypotheses testing, any step-up or step-down procedure is concordant.

*Theorem 3.* If the set of all the $p$-values is PRDS on the subset of $p$-values corresponding to true null hypotheses, the selection rule is concordant, and the procedure used for testing each selected family is

(a) the Bonferroni procedure or
(b) the BH procedure,

then the selection-adjusted procedure guarantees in case (a)

$$E\left[\sum_{i \in \mathcal{S}(\mathbf{P})} V_i \Big/ \max\{|\mathcal{S}(\mathbf{P})|, 1\}\right] \leqslant q$$

and in case (b)

$$E\left[\sum_{i \in \mathcal{S}(\mathbf{P})} \mathrm{FDP}_i \Big/ \max\{|\mathcal{S}(\mathbf{P})|, 1\}\right] \leqslant q.$$

The proof is given in Appendix D.

## 5. Selection of the families via multiple-hypotheses testing

As is often encountered in practice, the selected families may be considered scientific findings by themselves, so it would be appropriate to address the erroneous selection of a family, and to control some error rate of the selection process. So did Heller *et al.* (2009) for selecting gene sets in microarray analysis and Sun and Wei (2011) for analysing time course experiments. We may then associate each family with its global null (intersection) hypothesis that all we observe in that family is pure noise and combine the inside-family $p$-values to construct a valid $p$-value for this intersection hypothesis. (See Loughin (2004) for a systematic comparison of combining functions that can be used for this.) Using a multiple-testing procedure on these combined $p$-values, and selecting the families accordingly, is a natural approach.

The choice of the multiple-testing procedure should be guided by the error rate that we wish to control at the family level and the dependence among the combined $p$-values. For example, to achieve control of FDR at the family level (i.e. the expected proportion of all-null families among the selected families), we may apply the BH procedure on the combined $p$-values, if the combined $p$-values are PRDS on the subset of combined $p$-values corresponding to all-null families. For FWER-control at the family level (i.e. control of the probability that one or

more all-null families are selected) under the same dependence structure, we may use Hochberg's procedure, or revert to the Bonferroni procedure if the dependence is more general. Other testing procedures may be used as well: any step-up or step-down procedure defines a simple selection rule (see Appendix E) and can be used in procedure 1. For other procedures such as adaptive FDR- and FWER-procedures (Benjamini and Hochberg, 2000; Storey *et al*., 2004; Benjamini *et al*., 2006; Blanchard and Roquain, 2009; Sarkar *et al*., 2012) that are not simple procedure 2 can be used.

In some cases the control on the average of $\mathcal{C}$ by the selection-adjusted procedure yields control of FDR at the family level, thus offering control not only for the erroneous discoveries within each family but also at the first stage where the families are selected. This happens when the selection rule and the testing procedure are such that in each selected family at least one rejection is made, and the error measure $\mathcal{C}=1$ when one or more rejections are made in an all-null family (e.g. the error rate is FWER, FDX or FDR). To see this, note that, if $I_0 \subseteq \{1,\ldots,m\}$ is the set of indices of the all-null families and $I_1 \subseteq \{1,\ldots,m\}$ is the set of indices of the families containing one or more false null hypotheses,

$$E(\mathcal{C}_s) = E\left[\sum_{i\in\mathcal{S}(\mathbf{P})\cap I_0} \mathcal{C}_i \Big/ \max\{|\mathcal{S}(\mathbf{P})|,1\}\right] + E\left[\sum_{i\in\mathcal{S}(\mathbf{P})\cap I_1} \mathcal{C}_i \Big/ \max\{|\mathcal{S}(\mathbf{P})|,1\}\right].$$

From the above conditions it follows that, for each $i\in\mathcal{S}(\mathbf{P})\cap I_0$, $\mathcal{C}_i=1$; therefore

$$E\left[\sum_{i\in\mathcal{S}(\mathbf{P})\cap I_0} \mathcal{C}_i \Big/ \max\{|\mathcal{S}(\mathbf{P})|,1\}\right] = E[|\mathcal{S}(\mathbf{P})\cap I_0|/\max\{|\mathcal{S}(\mathbf{P})|,1\}],$$

which is FDR at the family level. It follows that if it is guaranteed that in each selected family at least one rejection is made, and the procedures used for testing the selected families are FWER, FDR or FDX controlling procedures, both control of $E(\mathcal{C}_{\mathcal{S}})$ and control of FDR at the family level are guaranteed by procedures 1 and 2, under the conditions of theorem 1 for the first, and under the conditions of theorems 2 and 3 for the latter.

We shall now show that for any collection of testing procedures used at step 2 of procedure 1 we can find a selection rule that guarantees that there is at least one rejection in each selected family. For that let $P_i^{\mathrm{adj}}$ be the minimum adjusted $p$-value within family $i$, according to the testing procedure that is used to control $E(\mathcal{C}_i)$ for $i=1,\ldots,m$. Select the families for which $P_i^{\mathrm{adj}} \leqslant Rq/m$, where $R$ is the number of selected families. It is easy to see that applying the BH procedure on the set of minimum adjusted $p$-values, $\{P_1^{\mathrm{adj}}, P_2^{\mathrm{adj}},\ldots,P_m^{\mathrm{adj}}\}$, and selecting the families for which the minimum adjusted $p$-value is rejected defines a selection rule that satisfies the above property, with the additional condition that the maximal such $R$ is chosen. Therefore, our recommendation is to use the BH procedure on the set of minimum adjusted $p$-values as the selection rule at step 1 of procedure 1.

Heller *et al*. (2009) addressed a similar problem of inference across families of hypotheses in microarray analysis. They first selected promising gene sets and then looked for differentially expressed genes within these gene sets. They defined an erroneous discovery of a set if a set is selected while no gene in the set is differentially expressed, or if it is appropriately selected but one of the genes in the set is erroneously discovered. They defined the overall false discovery rate criterion as the expected proportion of 'erroneous' discoveries of gene sets out of all the selected gene sets. This error criterion is equivalent to $E(\mathcal{C}_{\mathcal{S}})$ for $\mathcal{C}=\mathbf{I}_{\{V\geqslant 1\}}$ when the selection criterion and the procedure that is used for testing the selected families guarantee that in each family (gene set) at least one rejection is made. This condition is not always fulfilled. For example,

when the signal in the family is weak, it may be possible to see evidence that there is at least one signal in this family but impossible to point out where this signal is. In these cases our criterion does not coincide with the overall false discovery rate. To see it, suppose that an all-null family is selected, and there are no rejections inside this family. This family will have no contribution to $\mathcal{C}_S$, whereas it will have a contribution to the proportion of erroneous discoveries of gene sets out of all the selected gene sets, as defined by Heller *et al*. (2009).

In Heller *et al*. (2009) the division of the hypotheses into families was determined by the problem. In many applications each hypothesis carries two 'tags', i.e. the hypotheses have two-ways structure, where families can be constructed by pooling along either dimension. In these cases the researcher should define the families by the most important dimension for inference. In Section 7 we show an example of such an application.

## 6. Addressing the power of the method proposed

In this work we focus on the error criterion $E(\mathcal{C}_S)$ and show that it may be a relevant error measure in the applications. Although our goal is not to gain power and comparing the power of procedures that control different error rates is always questionable, we would like to sketch the issues that are relevant to such a comparison between a method controlling the expected average FDP and the BH procedure applied on the combined set of hypotheses controlling FDR.

We compare the power of the BH procedure applied on the combined set of hypotheses (the combined BH procedure) with that of procedure 1, where the combined $p$-values are constructed by using Simes's method, the BH procedure making use of these $p$-values is the selection rule and the BH procedure is used for testing the hypotheses in each family (denoted as the $(\mathrm{BH}-q, \mathrm{BH}-Rq/m)$ procedure). In the simulations the test statistics are normally distributed with standard deviation 1, mean 0 for the true null hypotheses and a common mean $\mu > 0$ for the false null hypotheses. In Fig. 1 we show the estimated average power as a function of the common expectation under the alternative hypothesis. We show the power under two settings: 1, nine all-null families containing 100 hypotheses and one family containing five hypotheses, where all the null hypotheses are false; 2, five all-null families of size 5, and five families containing 100 hypotheses where half of the hypotheses are false.

In setting 1 the power of the $(\mathrm{BH}-q, \mathrm{BH}-Rq/m)$ procedure is higher than the power of the combined BH procedure; see Fig. 1(a). In setting 2 the power of the $(\mathrm{BH}-q, \mathrm{BH}-Rq/m)$ procedure is lower than the power of the combined BH procedure; see Fig. 1(b). We demonstrate that we may gain or lose power by controlling the expected average FDP rather than the overall FDR. A gain in power is expected when there is homogeneity within the families, i.e. each family either consists almost only of true null hypotheses or consists almost only of false null hypotheses, and most of the false null hypotheses are clustered in small families.

## 7. Associating single-nucleotide polymorphisms with brain volume

We would like to show the relevance of our approach to voxelwise genomewide association study by using the study of Stein *et al*. (2010) that was described in Section 1. We first outline the analysis that they conducted and try to understand their concerns as reflected in the way that they presented and discussed their results. In view of that understanding, we propose to use the methodology that is presented in this paper. Needless to say that, as the approach that is presented in this paper is novel, in no way do we criticize the original analysis: it presented a very reasonable computational compromise while trying to address selective inference issues with the then available tools.

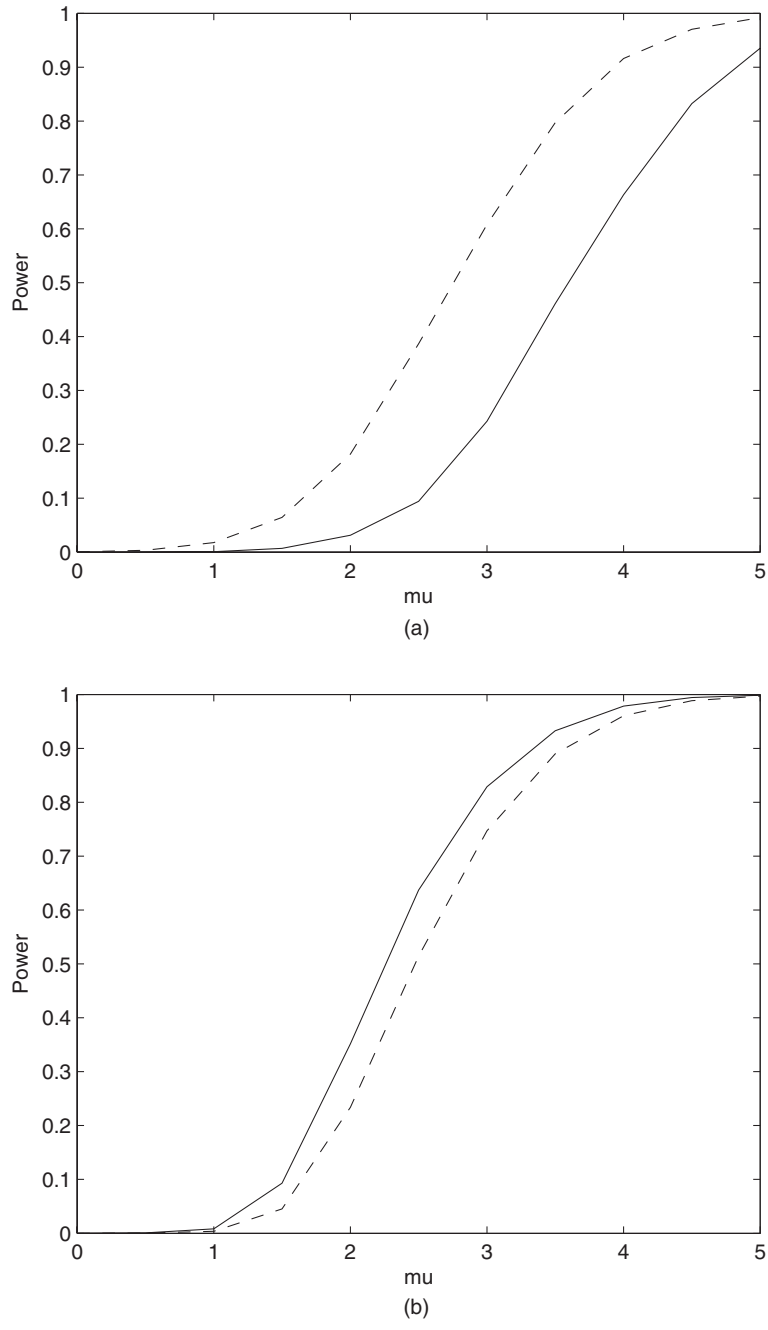


**Fig. 1.** Estimated average power of procedure 1 where the families are selected by using the BH procedure at level 0.05 on the Simes $p$-values for the family intersection hypotheses, and the selected families are tested by using the BH procedure at level $R/m \times 0.05$, where $R/m$ is the proportion of selected families (– – –), and the estimated average power of the BH procedure at level 0.05 applied on the combined set of $p$-values (———): (a) setting 1, nine all-null families containing 100 hypotheses and one family containing five hypotheses, where all the null hypotheses are false; (b) setting 2, five all-null families of size 5 and five families containing 100 hypotheses where half of the hypotheses are false

### *7.1. The analysis in Stein et al. (2010)*

Stein *et al*. (2010) explored the relationship between each of 448293 SNPs and each of 31622 voxels of the entire brain across 740 elderly subjects, including subjects with Alzheimer's disease, mild cognitive impairment and healthy elderly controls from the Alzheimer's disease neuroimaging initiative. The phenotype of interest was the percentage volume difference relative to a sample-specific template at each voxel, and a regression was conducted at each SNP with the phenotype as the dependent variable and the number of minor alleles, age and sex as the independent variables (assuming the additive genetic model). In the original analysis for each voxel only the most significantly associated SNP was retained for further use. Its $p$-value was adjusted to obtain uniform distribution when no SNP is associated with that voxel. The BH procedure was applied on the such-adjusted $p$-values. No SNP was found significant at the 0.05 level. Only two SNPs were found at the 0.5 level. Stein *et al*. (2010) were still interested in the most significant five SNPs, and for each of them the voxels for which the SNP was the most associated with were mapped.

### *7.2. Analysis using the control on the average*

According to the presentation of the results in Stein *et al*. (2010), they wished to find SNPs that are associated with some regions in the brain, and then to be able to make maps of these regions per each selected SNP. Adopting this goal, we define each family as the set of association hypotheses between all voxels and a specific SNP. This way, selection of the families would be equivalent to the selection of SNPs with at least one non-null association, which could be followed by searching the voxels that are associated with the selected SNPs.

The next practical question is what error rates should be controlled in this problem? It is obvious that the investigators do not wish to emphasize each voxel–SNP pair where an association is found; therefore there is no need to control for a global error rate, on the combined set of all the pairs discovered. The emphasis is on the selected SNPs and on the regions in the brain that could be affected by these SNPs. Therefore, it would be reasonable

(a) to control for some error rate when selecting the SNPs and
(b) to obtain some level of confidence within the selected SNPs by controlling the expected average measure of errors over the selected SNPs.

The choice of the procedure for testing across families should be guided by the desired error rate for the selected SNPs (see Section 5). Since the selected SNPs are proposed for further research, FDR-control at the level of SNPs seems appropriate. Theorems 1 and 2 offer the methods to obtain the control of erroneously identified voxels on the average over the selected SNPs. According to these theorems, any commonly used method in magnetic resonance imaging research could be applied across voxels for each selected SNP separately at the adjusted level: resampling or random-field theory FWER controlling approaches for the control of the expected number of SNPs with at least one erroneously discovered voxel out of all the selected SNPs, or the BH FDR controlling procedure for the control of expected average proportion of falsely discovered voxels over the selected SNPs. Theorems 1 and 2, however, assume independence across SNPs. Theorem 3 shows that the result remains true under a certain type of positive dependence (PRDS), if for each selected SNP the BH procedure is applied across voxels at the adjusted level. Since it is reasonable to assume positive dependence between the voxels and between the SNPs, the overall positive dependence across the $p$-values seems reasonable.

In view of this discussion, we suggest the following method.

*Step 1*: for each SNP, calculate the $p$-value for the association between the SNP and the volume at each voxel, controlling for age and sex.
*Step 2*: calculate the intersection hypothesis for each SNP by combining all the 31622 $p$-values calculated for that SNP at stage 1 by using Simes's test for the intersection hypothesis.
*Step 3*: test all the 448293 intersection hypotheses by using the BH procedure at level 0.05 by using the $p$-values calculated at stage 2. Let $R$ be the number of rejected intersection hypotheses.
*Step 4*: select the $R$ SNPs for which the intersection hypothesis was rejected at stage 3. For each selected SNP, apply the BH procedure at level $R \times 0.05/448293$ on all the 31622 $p$-values calculated at stage 1.

Assuming positive dependence between all the $p$-values, we expect that stages 1–3 guarantee FDR-control at the level of SNP families, on the basis of the simulation study in Benjamini and Heller (2008) regarding the application of the BH procedure on Simes's $p$-values under positive dependence. Moreover, according to theorem 3, we obtain that stages 1–4 guarantee control of the expected average false discovery proportion over the selected SNPs.

The original analysis can also be described by our current point of view: each family was tagged by the voxel—the hypotheses on the association of each SNP with a specific voxel—the corrected $p$-value for each voxel was the $p$-value for testing the family intersection hypothesis. If testing would continue within each selected voxel and calibrated appropriately we could obtain control on the average over the selected voxels. Hence selected voxels should be reported, each with its associated SNPs. In this study there is a clear preference to tag the families by SNP, which is the way that Stein *et al*. (2010) chose to present their results. In other cases involving more than one way of partitioning into families, there may be justification for either mode of analysis, depending on the interpretation needed.

### 7.3. Results

The $p$-values at stage 1 were calculated by using the Plink program. Whereas Stein *et al*. (2010) calculated the minimum $p$-value for each voxel, which results in 31622 $p$-values, we had to calculate all the $31622 \times 448293 \approx 14$ billion $p$-values. Since it was impossible to keep all the $p$-values, those larger than 0.1 were discarded but the fact that they were inspected was taken into consideration: their values were used as if they were increased to 1. At stage 2 the $p$-values for the intersection hypotheses were calculated as follows. Assume that $k$ is the number of $p$-values that are less than 0.1 for some SNP. Then we calculated $31622 \min_{j=1,\ldots,k}(P_{i(j)}/j)$, where $P_{i(j)}$ is the $j$th largest $p$-value in family $i$. This value will always be larger than or equal to Simes's $p$-value and is equivalent to replacing the values that are larger than 1 by 1, resulting in a valid $p$-value for the intersection hypothesis. Applying step 3 and step 4 on the basis of these $p$-values, we obtain the same results as if we used the original Simes $p$-values, since the hypotheses with $p$-values larger than 0.1 may not be rejected by the BH procedure at level 0.05.

The number of SNPs that were selected at stage 3 was 11. For each selected SNP the maps of associated voxels were produced by applying the BH procedure on the $p$-values testing the association of that SNP with each one of the voxels at level $11 \times 0.05/448293 = 1.2 \times 10^{-6}$. For one SNP 197 voxels were found associated, and for the other 10 fewer than 57 were found associated. The overall median was 12 voxels, the three smallest regions comprising two voxels each. We intend to investigate the biological meaning of the results in co-ordination with the biological team that was involved in the original analysis.

We emphasize that in this new analysis our goal is not the gain in power. The error rates controlled, both at the level of SNPs and at the level of voxels within the selected SNPs, assure

that the biological meaning of the results, in the way they are of interest to the biologists, is valid.

We may use these results to demonstrate the difference between control on the average over the selected and overall control of the error rate, and we argue that the overall FDR does not offer a meaningful interpretation of the results. We discovered 11 regions, where the largest region comprised 197 voxels and the three smallest regions comprised two voxels each. Assume that three erroneous associations were discovered. We would be more confident in the structure of the regions discovered if these errors were in the region comprising 197 voxels, rather than in the situation where there was one error in each region comprising two voxels. In the latter case we have a very low level of confidence for three out of 11 regions, whereas in the first case we have a high level of confidence for all the regions discovered. This difference is reflected in the error measure that we used: when the three errors are in the large region, the average FDP is $1.4 \times 10^{-3}$, whereas when there is one error in each region comprising two voxels the average FDP is much higher: $0.5 \times 3/11 = 0.136$. However, the difference between these two situations is not reflected in the overall FDP: it is below 0.015 in both cases.

## 8. Discussion

There have been very few works (outside Heller *et al.* (2009) that was discussed in Section 5) that address formally the issue of inference across families. We have mentioned Efron (2008) in Section 2. Other works dealing with this issue are Hu *et al.* (2010) and Sun and Wei (2011). Neither of these addressed the testing of multiple families of hypotheses within the framework of selective inference, which is the concern in our work. Testing each family separately while attending to some error rate control within each tested family has an obvious advantage that the control is achieved on the average across families. However, once only some families are selected on the basis of the same data, and inference is made or reported on only the selected families, even this simple average error rate across families deteriorates. In this paper we pointed at this danger, formulated it and offered simple—even if not optimal—ways to address it. In remark 4, we mentioned the ability to improve procedure 1 by estimating the number of families consisting of only false null hypotheses. The advantage of procedure 1 is that it offers an adjustment which is sufficient for the control of $E(\mathcal{C}_{\mathcal{S}})$ for any selection rule that is used. The investigator may even not know how exactly the families were selected; the required information is only the number of selected families and the fact that the selection rule is simple. However, for a given specific selection rule, there may be a less stringent adjustment offering the control of $E(\mathcal{C}_{\mathcal{S}})$.

Sometimes, the situation that is faced calls for more stringent control. This is so when interest lies in assuring simultaneous control of the error rate across families, and not merely on the average over the selected families. Such a concern for simultaneity of inference across selected families can be formulated by $E(\max_{i=1,\ldots,m} \mathcal{C}_i)$. For example, in the case $\mathcal{C}_i = \mathbf{I}_{\{\mathrm{FDP}_i > \gamma\}}$ this is the probability that in at least one family the false discovery proportion is greater than $\gamma$. It is easy to see that controlling $E(\mathcal{C}_i)$ at level $q/m$ in *each* family guarantees control of this error criterion. However, in current complex applications interest rarely lies in all the families, but rather in the promising families. Therefore we address only the selective goal in this work.

It may sometimes happen that there are no rejections in a selected family. For example, when the signal in the family is weak, it may be possible to have enough evidence that there is at least one signal in this family, but impossible to point out where the signal is. Some investigators may claim that in this case the interpretation of the results is not intuitive; therefore they wish

to have at least one rejection in each selected family. We showed that for any testing procedure that is used within the families, when using the BH procedure on the minimal adjusted $p$-values for the selection of families and then adjusting for the selection, it is guaranteed that in each selected family at least one rejection is made.

The framework of testing only the selected families is similar to the hierarchical testing framework that was developed in Yekutieli *et al*. (2006), who defined hierarchical testing of trees of hypotheses in the following way:

(a) each hypothesis is associated with a single parent hypothesis (except for the hypotheses at the first level of the tree);
(b) each family is tested only if its parent hypothesis is rejected (except for the hypotheses at the first level of the tree, that are always tested).

Yekutieli *et al*. (2006) defined different FDR-types on trees of hypotheses. Each FDR-type is actually the false discovery rate restricted to the set of hypotheses which are of interest to the investigator.

(i) Full tree FDR: interest lies in the entire set of rejected hypotheses.
(ii) Level $l$ FDR: hypotheses of interest are those residing at some specific level $l$, chosen in advance.
(iii) Outer nodes FDR: the set of hypotheses of interest are the rejected hypotheses which are not parents to other rejected hypotheses.

Yekutieli (2008) proved that, when all the $p$-values of the tree are jointly independent, the hierarchical testing procedure where each family is tested by using the BH procedure at a certain level $q_1 < q$ controls the full tree and outer nodes FDR at level $q$ but does not always control the level $l$ FDR.

From this hierarchical point of view, our structure is that of a two-level tree where the parent hypothesis for each family is its intersection hypothesis, and the level 2 FDR is the global FDR for all the discoveries within the families. Yet there are two major differences between the two methodologies.

(a) We gave up on global FDR in favour of control on the average over the selected.
(b) Unlike Yekutieli (2008), who assumed independence between the tests of the parent and the child nodes, we allow these tests to be highly dependent, as in the case of the parent being the intersection test of its child hypotheses.

In this paper we mainly addressed the goal of controlling the expected average measure of error over the selected families. Other types of error criteria may be relevant as well. The investigator might wish to control for some error rate on the pooled set of discoveries across all the families. This seems to be the only concern in Efron (2008) and Hu *et al*. (2010). If the selected families are considered as scientific findings by themselves, which is a situation that is often encountered in large testing problems, it would be appropriate to address the erroneous selection of a family, and to control some error rate of the selection process, as, for example suggested by Heller *et al*. (2009) in the context of microarray analysis and Sun and Wei (2011) in the context of time course experiments. The investigator may be interested in more than one type of error measures. For example, one might wish to control for FDR within each gene set, on the average over the gene sets selected and globally on the pooled set of genes discovered across all the gene sets. Therefore, an interesting research direction could be development of the procedures controlling concurrently several error measures that are of interest to the investigator.

## Acknowledgements

We thank Dr Stein and Professor Thompson for making the preprocessed data of the SNP study available to us, and Jonathan Rosenblatt for solving the difficult computational problems due to the enormous number of hypotheses tested. The research leading to these results has received funding from the European Research Council under EC–EP7 European Research Council grant PSARPS-297519, and from US Department of Defense grant W81XWH-11-2-008.

## Appendix A: Computation of the conditional probability for example 1

We consider a family where eight hypotheses are true null hypotheses with uniform $p$-values and two hypotheses are false null hypotheses, each having a $p$-value whose cumulative distribution function is the squared root of the uniform cumulative distribution function. Without loss of generality, let $P_1$ and $P_2$ be the $p$-values corresponding to false null hypotheses and $P_3, \ldots, P_{10}$ be the $p$-values corresponding to true null hypotheses. Let $P_{(1)}$ be the minimal $p$-value in the family and let $V$ be the number of type I errors in the family. Then the conditional probability that there is at least one type I error within the family is

$$\begin{aligned}
\Pr(V>0|P_{(1)}\leqslant 0.05/10) &= 1-\Pr(V=0|P_{(1)}\leqslant 0.05/10)\\
&= 1-\frac{\Pr[\{P_3,P_4,\ldots,P_{10}>0.05/10\}\cap(\cup_{j=1}^{2}\{P_j\leqslant 0.05/10\})]}{\Pr(P_{(1)}\leqslant 0.05/10)}\\
&= 1-\frac{[2\sqrt{(0.05/10)}\{1-\sqrt{(0.05/10)}\}+0.05/10](1-0.05/10)^8}{1-\{1-\sqrt{(0.05/10)}\}^2(1-0.05/10)^8}\\
&= 0.23.
\end{aligned}$$

## Appendix B: Formula for $E(\mathcal{C}_{\mathcal{S}})$ in example 2

In example 2, the formula for $E(\mathcal{C}_{\mathcal{S}})$ is

$$\left\{1-\left(1-\frac{q}{n}\right)^n\right\}\frac{1-(1-q)^{nm}}{1-(1-q)^n}. \tag{9}$$

The proof is as follows. In this case

$$E(\mathcal{C}_{\mathcal{S}})=\sum_{i=1}^{m}\sum_{k=1}^{m}\frac{1}{k}\Pr\{V_i>0, i\in\mathcal{S}(\mathbf{P}), C_k^{(i)}\}.$$

Family $i$ is selected if its minimal $p$-value is less than $q$. Each selected family is tested by using the Bonferroni procedure at level $q$. Since all the null hypotheses are true, there is at least one type I error in family $i$ if its minimal $p$-value is less than $q/n$. Therefore, if Bonferroni tested, each family where at least one type I error is made is selected. Now we obtain

$$\begin{aligned}
E(\mathcal{C}_{\mathcal{S}}) &= \sum_{i=1}^{m}\sum_{k=1}^{m}\frac{1}{k}\Pr(V_i>0, C_k^{(i)})\\
&= \sum_{i=1}^{m}\sum_{k=1}^{m}\frac{1}{k}\Pr(V_i>0)\Pr(C_k^{(i)})\\
&= \sum_{i=1}^{m}\left\{1-\left(1-\frac{q}{n}\right)^n\right\}\sum_{k=1}^{m}\frac{1}{k}\binom{m-1}{k-1}\{1-(1-q)^n\}^{k-1}(1-q)^{n(m-k)}.
\end{aligned}$$

Let us define random variable $Y\sim \mathrm{Bin}\{m-1, 1-(1-q)^n\}$. It is easy to see that

$$E(\mathcal{C}_{\mathcal{S}})=m\left\{1-\left(1-\frac{q}{n}\right)^n\right\}E\left(\frac{1}{Y+1}\right). \tag{10}$$

Using the proof of lemma 1 in Benjamini *et al.* (2006) we obtain

$$E\left(\frac{1}{Y+1}\right)=\frac{1-(1-q)^{nm}}{m\{1-(1-q)^{n}\}}. \tag{11}$$

Substituting equation (11) in equation (10) we obtain formula (9).

## Appendix C: Proof of theorem 2

For each error criterion $E(\mathcal{C})$, let $\mathcal{C}_+$ be the support of random variable $\mathcal{C}$. As in Benjamini and Yekutieli (2005), we define the following series of events:

$$C_k^{(i)} := \{\mathbf{P}^{(i)} : R_{\min}(\mathbf{P}^{(i)}) = k\}. \tag{12}$$

According to the definition of $R_{\min}(\mathbf{P}^{(i)})$ (see equation (8) in Section 3), for each value of $\mathbf{P}^{(i)}$ and $P_i = p_i$, such that $i \in \mathcal{S}(\mathbf{P}^{(i)}, p_i)$, $R_{\min}(\mathbf{P}^{(i)}) \leqslant |\mathcal{S}(\mathbf{P}^{(i)}, p_i)|$. Therefore,

$$\begin{aligned} E(\mathcal{C}_{\mathcal{S}}) &= E\left[\frac{\sum_{i\in\mathcal{S}(\mathbf{P})}\mathcal{C}_i}{\max\{|\mathcal{S}(\mathbf{P})|,1\}}\right] \\ &\leqslant \sum_{i=1}^{m}\sum_{k=1}^{m}\frac{1}{k}\sum_{c\in\mathcal{C}_+} c\Pr\{\mathcal{C}_i = c, i\in\mathcal{S}(\mathbf{P}), C_k^{(i)}\}. \end{aligned} \tag{13}$$

Expression (13) is identical to expression (4) in the proof of theorem 1, but the definition of $C_k^{(i)}$ here is different. In the proof of theorem 1 we use only the facts that the event $C_k^{(i)}$ is defined on the space $P^{(i)}$ and that, for each $i = 1, \ldots, m$, $\Sigma_{k=1}^{m}\Pr(C_k^{(i)}) = 1$. These facts remain true for the series of events defined in expression (12). Therefore, the arguments that were used in the proof of theorem 1 after obtaining equation (4) can be applied here.

## Appendix D: Proof of theorem 3

The proof of theorem 3 uses the techniques that were developed in Benjamini and Yekutieli (2001, 2005). The proof in case (b) is much more involved than in case (a).

### *D.1. Proof for case (a)*

For each $i = 1, \ldots, m$, let $m_i$ be the number of hypotheses in family $i$ and $m_{0i}$ be the number of true null hypotheses in family $i$. Let $H_{0ij}$ and $P_{ij}$, $j = 1, \ldots, m_i$, be the hypotheses and the $p$-values in family $i$, $i = 1, \ldots, m$. We shall use the series of events $C_k^{(i)}$, which were defined in Appendix C, expression (12).

We shall prove that

$$\begin{aligned} E\left[\frac{\sum_{i=1}^{m}V_i}{\max\{|\mathcal{S}(\mathbf{P})|,1\}}\right] &= \sum_{i=1}^{m}\sum_{k=1}^{m}\sum_{j=1}^{m_{0i}}\frac{1}{k}\Pr\left\{i\in\mathcal{S}(\mathbf{P}), C_k^{(i)}, P_{ij}\leqslant\frac{kq}{mm_i}\right\} \\ &\leqslant \sum_{i=1}^{m}\sum_{k=1}^{m}\sum_{j=1}^{m_{0i}}\frac{1}{k}\Pr\left(C_k^{(i)}, P_{ij}\leqslant\frac{kq}{mm_i}\right) \qquad (14) \\ &= \sum_{i=1}^{m}\sum_{k=1}^{m}\sum_{j=1}^{m_{0i}}\frac{1}{k}\Pr\left(C_k^{(i)}|P_{ij}\leqslant\frac{kq}{mm_i}\right)\Pr\left(P_{ij}\leqslant\frac{kq}{mm_i}\right) \\ &\leqslant \sum_{i=1}^{m}\sum_{k=1}^{m}\sum_{j=1}^{m_{0i}}\frac{1}{k}\Pr\left(C_k^{(i)}|P_{ij}\leqslant\frac{kq}{mm_i}\right)\frac{kq}{mm_i} \qquad (15) \\ &= \frac{q}{m}\sum_{i=1}^{m}\frac{1}{m_i}\sum_{j=1}^{m_{0i}}\sum_{k=1}^{m}\Pr\left(C_k^{(i)}|P_{ij}\leqslant\frac{kq}{mm_i}\right). \qquad (16) \end{aligned}$$

Inequality (14) is obtained by dropping the condition $i \in \mathcal{S}(\mathbf{P})$. Inequality (15) is true since the $p$-values corresponding to true null hypotheses have a uniform (or stochastically larger) distribution. We shall now prove that, for any $i = 1, \ldots, m$ and $j = 1, \ldots, m_{0i}$,

$$\sum_{k=1}^{m} \Pr\left( C_k^{(i)} | P_{ij} \leqslant \frac{kq}{mm_i} \right) \leqslant 1. \tag{17}$$

Since the selection rule is concordant, the set $D_k^{(i)} = \cup_{j=1}^{k} C_j^{(i)}$, which can be written as $\{P^{(i)} : R_{\min}(P^{(i)}) < k+1\}$, is an increasing set. The PRDS property on the subset of the $p$-values corresponding to the true null hypotheses implies that, for any $i = 1, \ldots, m$ and $j = 1 \ldots, m_{0i}$,

$$\Pr(D_k^{(i)} | P_{ij} \leqslant \alpha) \leqslant \Pr(D_k^{(i)} | P_{ij} \leqslant \alpha')$$

for any $\alpha \leqslant \alpha'$. Now, we obtain for any $k = 1, \ldots, m-1$

$$\begin{aligned} \Pr\left( D_k^{(i)} | P_{ij} \leqslant \frac{kq}{mm_i} \right) + \Pr\left\{ C_{k+1}^{(i)} | P_{ij} \leqslant \frac{(k+1)q}{mm_i} \right\} & \\ \leqslant \Pr\left\{ D_k^{(i)} | P_{ij} \leqslant \frac{(k+1)q}{mm_i} \right\} + \Pr\left\{ C_{k+1}^{(i)} | P_{ij} \leqslant \frac{(k+1)q}{mm_i} \right\} & \\ = \Pr\left\{ D_{k+1}^{(i)} | P_{ij} \leqslant \frac{(k+1)q}{mm_i} \right\}. & \end{aligned} \tag{18}$$

Applying repeatedly inequality (18) for $k = 1, \ldots, m-1$, and using the fact that $C_1^{(i)} = D_1^{(i)}$, we obtain

$$\sum_{k=1}^{m} \Pr\left( C_k^{(i)} | P_{ij} \leqslant \frac{kq}{mm_i} \right) \leqslant \Pr\left( D_m^{(i)} | P_{ij} \leqslant \frac{mq}{mm_i} \right) \leqslant 1. \tag{19}$$

Using expressions (16) and (19) we obtain

$$E\left[ \frac{\sum_{i=1}^{m} V_i}{\max\{|\mathcal{S}(\mathbf{P})|, 1\}} \right] \leqslant \frac{q}{m} \sum_{i=1}^{m} \frac{m_{0i}}{m_i} \leqslant q.$$

### *D.2. Proof for case (b)*

Let $P_i$ be the set of $p$-values corresponding to family $i$. For each $j = 1, \ldots, m_{0i}$, let $P_i^{(ij)}$ denote the set of the remaining $m_i - 1$ $p$-values after dropping $P_{ij}$. Let us define the following series of events on the range of $P_i^{(ij)}$. For family $i$, let $B_{r_i}^{(ij)}[k]$ denote the event in which, if $H_{0ij}$ is rejected by the BH procedure at level $kq/m$, $r_i$ hypotheses (including $H_{0ij}$) are rejected alongside it. We shall use again the series of events $C_k^{(i)}$, which was defined in Appendix C, expression (12). Using the fact that the $p$-values corresponding to the true null hypotheses have a uniform (or stochastically larger) distribution, we obtain

$$\begin{aligned} E\left( \frac{\sum_{i \in \mathcal{S}} \mathrm{FDP}_i}{\max\{|\mathcal{S}(\mathbf{P})|, 1\}} \right) &= \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \sum_{r_i=1}^{m_i} \frac{1}{r_i} \sum_{j=1}^{m_{0i}} \Pr\left\{ i \in \mathcal{S}(\mathbf{P}), C_k^{(i)}, B_{r_i}^{(ij)}[k], P_{ij} \leqslant \frac{r_i kq}{mm_i} \right\} \\ &\leqslant \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \sum_{r_i=1}^{m_i} \frac{1}{r_i} \sum_{j=1}^{m_{0i}} \Pr\left( C_k^{(i)}, P_{ij} \leqslant \frac{r_i kq}{mm_i}, B_{r_i}^{(ij)}[k] \right) \\ &\leqslant \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{1}{k} \sum_{r_i=1}^{m_i} \frac{1}{r_i} \sum_{j=1}^{m_{0i}} \frac{r_i kq}{mm_i} \Pr\left( C_k^{(i)}, B_{r_i}^{(ij)}[k] | P_{ij} \leqslant \frac{r_i kq}{mm_i} \right) \\ &= \frac{q}{m} \sum_{i=1}^{m} \frac{1}{m_i} \sum_{j=1}^{m_{0i}} \sum_{k=1}^{m} \sum_{r_i=1}^{m_i} \Pr\left( C_k^{(i)}, B_{r_i}^{(ij)}[k] | P_{ij} \leqslant \frac{r_i kq}{mm_i} \right). \end{aligned}$$

Now the fact that $m_{0i} \leqslant m_i$ reduces case (b) to the inequality

$$\sum_{k=1}^{m} \sum_{r_i=1}^{m_i} \Pr\left( C_k^{(i)}, B_{r_i}^{(ij)}[k] | P_{ij} \leqslant \frac{r_i kq}{mm_i} \right) \leqslant 1 \tag{20}$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, m_{0i}$.

For each $i = 1, \ldots, m$ and $t = 1, \ldots, mm_i$, let us define the group

$$I_t = \{(a, b) : a \in \{1, \ldots, m\}, b \in \{1, \ldots, m_i\}, ab = t\}.$$

Obviously, $I_t$ is a finite set. Note that

$$\sum_{k=1}^{m}\sum_{r_i=1}^{m_i}\Pr\left(C_k^{(i)}, B_{r_i}^{(ij)}[k]|P_{ij}\leqslant\frac{r_i kq}{mm_i}\right)=\sum_{t=1}^{mm_i}\sum_{(k,\,r_i)\in I_t}\Pr\left(C_k^{(i)}, B_{r_i}^{(ij)}[k]|P_{ij}\leqslant\frac{tq}{mm_i}\right).$$

Therefore, inequality (20) can be written in the form

$$\sum_{t=1}^{mm_i}\sum_{(k,\,r_i)\in I_t}\Pr\left(C_k^{(i)}, B_{r_i}^{(ij)}[k]|P_{ij}\leqslant\frac{tq}{mm_i}\right)\leqslant 1. \tag{21}$$

For each family $i$ and its hypothesis $H_{0ij}$, $i=1,\ldots,m$, $j=1,\ldots,m_{0i}$, let us define

$$A_s^{(ij)}\triangleq\bigcup_{t=1}^{s}\bigcup_{(k,\,r_i)\in I_t}\left(C_k^{(i)}\bigcap B_{r_i}^{(ij)}[k]\right).$$

The key statement to prove inequality (21) is the following proposition.

*Proposition 1.* For any $s=1,\ldots,mm_i-1$, $i=1,\ldots,m$, $j=1,\ldots,m_{0i}$, we have the inequality

$$\Pr\left(A_s^{(ij)}|P_{ij}\leqslant\frac{sq}{mm_i}\right)+\sum_{(k,\,r_i)\in I_{s+1}}\Pr\left\{C_k^{(i)}, B_{r_i}^{(ij)}[k]|P_{ij}\leqslant\frac{(s+1)q}{mm_i}\right\}\leqslant\Pr\left\{A_{s+1}^{(ij)}|P_{ij}\leqslant\frac{(s+1)q}{mm_i}\right\}. \tag{22}$$

We show that this proposition implies inequality (21). Note that

$$\Pr\left(A_1^{(ij)}|P_{ij}\leqslant\frac{q}{mm_i}\right)=\sum_{(k,\,r_i)\in I_1}\Pr\left(C_k^{(i)}, B_{r_i}^{(ij)}[k]|P_{ij}\leqslant\frac{q}{mm_i}\right).$$

Now, a consequent application of inequality (22) with $s=1,\ldots,mm_i-1$ leads to the inequality

$$\sum_{t=1}^{mm_i}\sum_{(k,\,r_i)\in I_t}\Pr\left(C_k^{(i)}, B_{r_i}^{(ij)}[k]|P_{ij}\leqslant\frac{tq}{mm_i}\right)\leqslant\Pr(A_{mm_i}^{(ij)}|P_{ij}\leqslant q)\leqslant 1,$$

which implies inequality (21).

*D.2.1. Proof of proposition 1*

We show that $A_s^{(ij)}$ is an increasing set. Let

$$D_k^{(i)}\triangleq\{P^{(i)}: R_{\min}(P^{(i)})<k+1\}$$

and

$$G_{r_i}^{(ij)}[k]\triangleq\left\{P_i^{(ij)}: P_{i(r_i)}^{(ij)}>\frac{(r_i+1)kq}{mm_i}, P_{i(r_i+1)}^{(ij)}>\frac{(r_i+2)kq}{mm_i},\ldots,P_{i(m_i-1)}^{(ij)}>\frac{kq}{m}\right\},$$

where $\{P_{i(1)}^{(ij)}\leqslant P_{i(2)}^{(ij)}\leqslant\ldots\leqslant P_{i(m_i-1)}^{(ij)}\}$ is the ordered set of $p$-values in the range of $P_i^{(ij)}$. It is easy to see that

$$A_s^{(ij)}=\bigcup_{t=1}^{s}\bigcup_{(k,\,r_i)\in I_t}\left(D_k^{(i)}\bigcap G_{r_i}^{(ij)}[k]\right).$$

Obviously, both $D_k^{(i)}$ and $G_{r_i}^{(ij)}[k]$ are increasing sets. Unions and intersections of increasing sets are also an increasing set; hence $A_s^{(ij)}$ is an increasing set. The PRDS property on the subset of $p$-values corresponding to the true null hypotheses implies that, for each $i=1,\ldots,m$, $j=1,\ldots,m_{0i}$ and any $\alpha\leqslant\alpha'$,

$$\Pr(A_s^{(ij)}|P_{ij}\leqslant\alpha)\leqslant\Pr(A_s^{(ij)}|P_{ij}\leqslant\alpha'). \tag{23}$$

It is obvious that, for $(k,r)\neq(k',r')$, the sets $(C_k^{(i)}\cap B_{r_i}^{(ij)}[k])$ and $(C_{k'}^{(i)}\cap B_{r_{i'}}^{(ij)}[k'])$ are disjoint. Therefore, for $t\neq t'$, $\cup_{(k,\,r_i)\in I_t}(C_k^{(i)}\cap B_{r_i}^{(ij)}[k])$ and $\cup_{(k,\,r_i)\in I_{t'}}(C_k^{(i)}\cap B_{r_i}^{(ij)}[k])$ are disjoint as well. Now we obtain for any $\alpha$

$$\Pr(A_s^{(ij)}|P_{ij}\leqslant\alpha)=\sum_{t=1}^{s}\sum_{(k,r_i)\in I_t}\Pr(C_k^{(i)}, B_{r_i}^{(ij)}[k]|P_{ij}\leqslant\alpha). \tag{24}$$

Using expressions (23) and (24) we obtain that, for any $s=1,\ldots,mm_i-1$,

$$\Pr\left(A_s^{(ij)}|P_{ij}\leqslant\frac{sq}{mm_i}\right)+\sum_{(k,\,r_i)\in I_{s+1}}\Pr\left\{C_k^{(i)},B_{r_i}^{(ij)}[k]|P_{ij}\leqslant\frac{(s+1)q}{mm_i}\right\}$$
$$\leqslant\Pr\left\{A_s^{(ij)}|P_{ij}\leqslant\frac{(s+1)q}{mm_i}\right\}+\sum_{(k,\,r_i)\in I_{s+1}}\Pr\left\{C_k^{(i)},B_{r_i}^{(ij)}[k]|P_{ij}\leqslant\frac{(s+1)q}{mm_i}\right\}$$
$$=\Pr\left\{A_{s+1}^{(ij)}|P_{ij}\leqslant\frac{(s+1)q}{mm_i}\right\}$$

and proposition 1 follows.

## Appendix E: Proof for the fact that step-up and step-down procedures define simple selection rules

Let $\alpha_1,\alpha_2,\ldots,\alpha_m$ be the critical values of the given procedure. Let $H_{0i}$ be a certain rejected hypothesis and $P_i$ be its $p$-value. We need to show that, when all the $p$-values excluding $P_i$ are fixed and $P_i$ changes as long as $H_{0i}$ is rejected, the total number of rejections remains unchanged.

Assume that this is a step-up procedure. Let $p_{(1)}^{(i)}\leqslant\ldots\leqslant\ldots\leqslant p_{(m-1)}^{(i)}$ be the ordered set of $p$-values excluding $P_i$. If $p_{(k-1)}^{(i)}\leqslant\alpha_k$, $p_{(k)}^{(i)}>\alpha_{k+1},\ldots,p_{(m-1)}^{(i)}>\alpha_m$, the number of rejections is $k$ for any value of $P_i$ which guarantees that $H_{0i}$ is rejected, i.e. $P_i\leqslant\alpha_k$.

Now assume that this is a step-down procedure. Let $p_{(1)}\leqslant p_{(2)}\leqslant\ldots\leqslant p_{(m)}$ be the ordered set of $p$-values. Assume that the number of rejections is $k$, thereby implying that $p_{(1)}\leqslant\alpha_1,\ldots,p_{(k)}\leqslant\alpha_k$, $p_{(k+1)}>\alpha_{k+1}$. Since $H_{0i}$ is rejected, $P_i=p_{(j)}$ for some $j\leqslant k$. Let us fix all the $p$-values excluding $P_i$ and change the value of $P_i$ so that $H_{0i}$ is still rejected. Assume that $\tilde{p}_{(1)}\leqslant\tilde{p}_{(2)}\leqslant\ldots\leqslant\tilde{p}_{(m)}$ is the ordered sequence of $p$-values after the value of $P_i$ has been changed. Now $P_i=\tilde{p}_{(j')}$ and, since $H_{0i}$ is rejected, $\tilde{p}_{(s)}\leqslant\alpha_s$ for each $s\leqslant j'$. If $j'=j$, it is obvious that the number of rejections remains unchanged. We shall now deal separately with two cases: $j'<j$ and $j'>j$.

(a) Assume that $j'<j$. Then $j'<k$; therefore it remains to show that $\tilde{p}_{(s)}\leqslant\alpha_s$ for $s=j'+1,\ldots,k$ and $\tilde{p}_{(k+1)}>\alpha_{k+1}$. Note that $\tilde{p}_{(s)}=p_{(s-1)}\leqslant\alpha_{s-1}\leqslant\alpha_s$ for $s=j'+1,\ldots,j$. For $s>j$, $\tilde{p}_{(s)}=p_{(s)}$; therefore now it is obvious that the number of rejections remains unchanged.

(b) Assume that $j'>j$. We shall now show that $j'\leqslant k$. Assume that $j\leqslant k<j'$. Note that, for each $j\leqslant s<j'$, $\tilde{p}_{(s)}=p_{(s+1)}$. Particularly, $\tilde{p}_k=p_{(k+1)}>\alpha_{k+1}\geqslant\alpha_k$, contradicting the rejection of $H_{0i}$. After we have proved that $j'\leqslant k$, the result follows immediately, since $\tilde{p}_{(s)}=p_{(s)}$ for $s>j'$.

## References


Benjamini, Y. and Heller, R. (2007) False discovery rate for spatial data. *J. Am. Statist. Ass.*, **102**, 1272–1281.

Benjamini, Y. and Heller, R. (2008) Screening for partial conjunction hypotheses. *Biometrics*, **64**, 1215–1222.

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

Benjamini, Y. and Hochberg, Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, **25**, 60–83.

Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2006) Adaptive linear step-up false discovery rate controlling procedures. *Biometrika*, **93**, 491–507.

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

Benjamini, Y. and Yekutieli, D. (2005) False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Am. Statist. Ass.*, **100**, 71–93.

Blanchard, G. and Roquain, E. (2009) Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.*, **10**, 2837–2871.

Efron, B. (2008) Simultaneous inference: when should hypotheses testing problems be combined? *Ann. Appl. Statist.*, **2**, 197–223.

Efron, B. and Tibshirani, R. (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidem.*, **23**, 70–86.

Farcomeni, A. (2008) A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statist. Meth. Med. Res.*, **17**, 347–388.

Genovese, C. R. and Wasserman, L. (2006) Exceedance control of the false discovery proportion. *J. Am. Statist. Ass.*, **101**, 1408–1417.

Heller, R., Manduchi, E., Grant, G. R. and Ewens, W. J. (2009) A flexible two-stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics*, **25**, 1019–1205.

Hu, J. X., Zhao, H. Y. and Zhou, H. H. (2010) False Discovery rate control with groups. *J. Am. Statist. Ass.*, **105**, 1215–1227.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004) Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statist. Appl. Genet. Molec. Biol.*, **3**, article 1042.

Lehmann, E. L. and Romano, J. P. (2005) Generalizations of the familywise error rate. *Ann. Statist.*, **33**, 1138–1154.

Loughin, T. (2004) A systematic comparison of methods for combining p-values from independent tests. *Computnl Statist. Data Anal.*, **47**, 467–485.

Pacifico, M. P., Genovese, C., Verdinelli, I. and Wasserman, L. (2004) False discovery control for random fields. *J. Multiv. Anal.*, **98**, 1441–1469.

Sarkar, S. K. (2007) Step-up procedures controlling generalized FWER and generalized FDR. *Ann. Statist.*, **35**, 2405–2420.

Sarkar, S. K., Guo, W. and Finner, H. (2012) On adaptive procedures controlling the familywise error rate. *J. Statist. Planng Inf.*, **142**, 65–78.

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M. J., Craig, D. W., Gerber, J. D., Allen, A. N., Corneveaux, J. J., DeChairo, B. M., Potkin, S. G., Weiner, M. W. and Thompson, P. M. (2010) Voxelwise genome-wide association study (vG-WAS). *NeuroImage*, **53**, 1160–1174.

Storey, J. D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, **31**, 2013–2035.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B, **66**, 187–205.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. and Mesirov, J. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natn. Acad. Sci. USA*, **102**, 15545–15550.

Sun, W. and Wei, Z. (2011) Multiple testing for pattern identification, with applications to microarray time-course experiments. *J. Am. Statist. Ass.*, **106**, 73–88.

Yekutieli, D. (2008) Hierarchical false discovery rate-controlling methodology. *J. Am. Statist. Ass.*, **103**, 309–316.

Yekutieli, D., Reiner-Benaim, A., Benjamini, Y., Elmer, G. I., Kafkafi, N., Letwin, N. E. and Lee, N. H. (2006) Approaches to multiplicity issues in complex research in microarray analysis. *Statist. Neerland.*, **60**, 414–437.