*Biometrika* (2021), **108**, 3, *pp.* 575–590
*Printed in Great Britain*
doi: 10.1093/biomet/asaa086
Advance Access publication 30 December 2020

# Hypotheses on a tree: new error rates and testing strategies

By MARINA BOGOMOLOV

*The William Davidson Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, Technion City, Haifa 3200003, Israel*
marinabo@technion.ac.il

CHRISTINE B. PETERSON

*Department of Biostatistics, Division of Basic Science Research, The University of Texas, MD Anderson Cancer Center, Houston, Texas 77030 , U.S.A.*
cbpeterson@mdanderson.org

YOAV BENJAMINI

*Department of Statistics and Operations Research, Tel-Aviv University, P.O. Box 39040, Tel-Aviv 6997801, Israel*
ybenja@tauex.tau.ac.il

and CHIARA SABATTI

*Department of Statistics, Stanford University, 50 Governor's Lane, Stanford, California 94305, U.S.A.*
sabatti@stanford.edu

## Summary

We introduce a multiple testing procedure that controls global error rates at multiple levels of resolution. Conceptually, we frame this problem as the selection of hypotheses that are organized hierarchically in a tree structure. We describe a fast algorithm and prove that it controls relevant error rates given certain assumptions on the dependence between the $p$-values. Through simulations, we demonstrate that the proposed procedure provides the desired guarantees under a range of dependency structures and that it has the potential to gain power over alternative methods. Finally, we apply the method to studies on the genetic regulation of gene expression across multiple tissues and on the relation between the gut microbiome and colorectal cancer.

*Some key words*: False discovery rate; Hierarchical testing; Multiple testing; Selective inference.

## 1. Introduction

In modern scientific studies researchers may be interested in examining a massive number of variables. Given the resulting large number of hypotheses considered, it is of critical importance to properly account for multiplicity, as this is one way to increase the replicability of results. Addressing this challenge, Benjamini & Hochberg (1995) proposed the false discovery rate, the expected proportion of rejected hypotheses that represent incorrect rejections, as a relevant

measure of error. The procedure to control the false discovery rate that they introduced has been widely applied across many fields; but despite its importance and success, it is not without limitations. In particular, it treats the entire collection of tested hypotheses as a single family; while this is appropriate in many circumstances, it overlooks structure that might be relevant for scientific investigations. Specifically, hypotheses may naturally be organized into classes or placed within a hierarchy. For example, in genomic studies, mutations can be grouped by the genes in which they occur, and studies often emphasize findings at this level, rather than focusing on individual genetic variants. Critically, controlling the false discovery rate in the entire collection of hypotheses, spanning all levels, does not guarantee error control for discoveries within a level. Conversely, controlling the false discovery rate only among the finer, more specific, hypotheses may not control error rates for discoveries grouped at a higher level. To demonstrate this limitation in particular, consider a case where many true discoveries are made in a rightfully discovered group, for example corresponding to many genetic variants within the same gene. The small proportion of false discoveries allowed by false discovery rate control may be scattered within many groups, as genes, that include no other true discovery, leading to a high group false discovery rate. Hence, even though the false discovery rate is controlled at the finer level, it is not controlled at the group level.

In this paper, we introduce a novel framework for testing hypotheses organized into a hierarchy with arbitrarily many levels. We propose both a new error rate and a testing procedure. Together, they guarantee control of error rates at multiple levels of resolution, with a built-in coordination between discoveries at different levels. To better appreciate the issues at hand, it is useful to survey previous work.

The presence of a hierarchical structure between the tested hypotheses has been recognized as being important for two reasons: (i) it allows researchers to control error rates relative to the discoveries that are finally reported, and (ii) it can be leveraged to increase power. With regard to (i), a number of papers have studied how to control the false discovery rate at multiple levels of resolution in specific contexts, such as fMRI imaging (Perone Pacifico et al., 2004; Benjamini & Heller, 2007), copy number variant detection (Siegmund et al., 2011) and genome-wide association studies (Brzyski et al., 2017). More generally, Benjamini & Bogomolov (2014) and Heller et al. (2017) considered the question of how to test groups of hypotheses at a high resolution when their $p$-values have been used to select promising groups at a low resolution. To this end, Benjamini & Bogomolov (2014) defined a version of the false discovery rate that incorporates the results of selection and introduced a procedure to control it. Heller et al. (2017), taking a different direction, proposed controlling conditional error rates by applying a multiple testing procedure within each selected group of hypotheses. This strategy can be pursued only when $p$-values conditioned on the selection of the group can be computed; see also Heller et al. (2019). Finally, the $p$-filter method of Foygel Barber & Ramdas (2016) controls the false discovery rate at the group level for each given division of the hypotheses into groups.

Explicitly accounting for the structure among hypotheses at the testing stage presents an opportunity to improve power. Exploring this idea, Yekutieli et al. (2006) and Yekutieli (2008) described a setting where hypotheses can be arranged in a tree structure, with the most specific or finest-resolution hypotheses in the leaf nodes. By testing the hypotheses in sequential order starting from the root of the tree, interesting branches can be identified, which would allow a more generous significance threshold for discoveries in those portions of the tree, while investigation into branches that contain low amounts of signal could be avoided, reducing the effort involved.

Another body of work dealing with trees and more general directed acyclic graphs includes the papers of Goeman & Mansmann (2008), Meinshausen (2008) and Rosenbaum (2008). Their procedures attempt to control the familywise error rate, or the probability of making any false

discoveries at all. While a strategy that limits the probability of at least one false discovery lends itself to easier interpretation across resolutions, controlling the familywise error rate can be overly stringent for studies focused on discovery or hypothesis generation and often results in low power. Previous work aimed at false discovery rate control in these settings either relies on very strong independence assumptions (Yekutieli, 2008) or proposes strategies that lead to the control of the false discovery rate on the total discoveries, without enabling interpretation of results at multiple layers of resolution (Lynch & Guo, 2016; Lei et al., 2020; Ramdas et al., 2019).

The present paper capitalizes on the results of Benjamini & Bogomolov (2014). Our procedure controls marginal error rates at each resolution, while making use of the hierarchical structure to potentially gain power. By testing from the most general to the more specific hypotheses down the hierarchy, our method selectively chooses which family of hypotheses to test. This approach is also adapted to sequential testing, as we do not require data for all the hypotheses upfront. The methodological development presented in this paper not only provides new theoretical results, but also allows us to consider a wide range of new applications, such as analysis of microbiome data, which follow a multi-level tree structure.

## 2. A tree of families of hypotheses

We consider a collection of hypotheses $\mathcal{F} = \{H_1, \ldots, H_m\}$ that are organized in a tree structure with $L$ levels, where hypotheses on the same level correspond to scientific statements made at the same resolution. In a microbiome study, for example, the root node might correspond to the hypothesis that there is no relation at all between bacterial communities in the oral cavity and the incidence of caries; the hypotheses at level 1 might specialize this statement to each of the phyla of bacteria, those in level 2 to each of the classes, and so on, following the branching of the taxonomic tree.

Formally, each hypothesis $H_i$ has only one parent, but can have multiple children, so that when one null hypothesis is true, all of its descendants are true, and when one null hypothesis is false, all of its ancestors are false. These logical relations hold, for example, when each parent hypothesis is the intersection of all its children; see Fig. 1 for an example. The level of a hypothesis corresponds to its distance from the root node. Because of the important role of levels, we explicitly include them in our notation, denoting by $\mathcal{F}_i^{\ell+1}$ the family of hypotheses at level $\ell + 1$ that has $H_i$ at level $\ell$ as its parent hypothesis.

The testing strategy that we are interested in is hierarchical, starting from the coarser hypotheses and increasing in resolution, similar to the procedure in Yekutieli (2008). The root node hypothesis $H_0$, residing at level 0, is regarded as a parent of the hypotheses at level 1. If there is no real interest in testing the hypothesis corresponding to this global null, we artificially add a root node and consider it to be always rejected. As we move down the tree, hypotheses are tested only if their parent node was rejected, and all hypotheses at the same level are tested at the same time.

Figure 1 depicts the organization of hypotheses and the direction of testing within a four-level tree. Ancestors of a given hypothesis are the hypotheses at the preceding levels that are connected to the given hypothesis by a path consisting of one or more arrows. For example, the ancestors of $H_4$ are $H_0$ and $H_1$, while the ancestors of $H_9$ are $H_0$, $H_1$ and $H_4$.

We assume that valid $p$-values for all the hypotheses in the tree can be calculated on the basis of the data, which can be gathered in an on-line fashion, with more precise information becoming available as we move down the tree, or prior to any testing. When the same data are used to test all hypotheses, the $p$-values across the tree will have complex dependencies, making it impossible to rely on the results in Yekutieli (2008). Scientists might use a variety of strategies to calculate $p$-values for hypotheses at different levels to maximize power, or they might rely on the $p$-values for

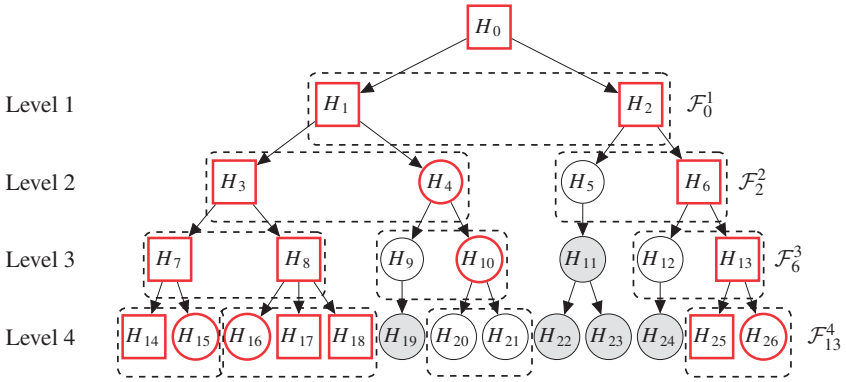


Fig. 1. Hierarchical structure of hypotheses in a four-level tree. Circles represent true null hypotheses, while squares denote false nulls. Children of the same parent constitute a family of hypotheses. To give an example of the sequential order of testing, nodes corresponding to tested hypotheses are unfilled, while grey nodes indicate hypotheses that are not tested. A red border distinguishes rejected hypotheses. Tested families are enclosed within dashed borders, with some labelled as $\mathcal{F}_i^j$ to illustrate the notation.

the finer-scale hypotheses and obtain the remainder with combination rules. The latter situation, for example, is typical of studies of genetic regulation, where $p$-values for the hypotheses of no association between individual variants and the expression of specific genes are calculated and then combined to test more global hypotheses, leading for instance to the discovery of genes whose expression is regulated by DNA variants. It is not the goal of this work to investigate the most powerful tests; rather, we focus on the development of multiplicity adjustment strategies that accept any collection of valid $p$-values for the hypotheses in the tree.

In our theory and simulations, aiming at a general-purpose rule, we emphasize the case where each parent hypothesis is the intersection of all its children, and the $p$-values for the level-$(\ell-1)$ hypotheses are derived by using the method of Simes (1986) on the $p$-values of the family of hypotheses they index at level $\ell$, starting from the available valid level-$L$ $p$-values and working up from the bottom of the tree. Specifically, the $p$-value for hypothesis $H_i$ at level $\ell-1$, indexing family $\mathcal{F}_i^\ell$, is obtained as

$$p = \min_j p_{(j)} \times \frac{k}{j}, \tag{1}$$

where the set $\{p_j : j = 1, \ldots, k\}$ corresponds to the collection of $p$-values for the hypotheses belonging to $\mathcal{F}_i^\ell$, so that $k = |\mathcal{F}_i^\ell|$, and the index in parentheses signifies that the $p$-values have been sorted in increasing order. Simes' rule is relatively robust to dependence, and has nice properties when used in conjunction with the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995); see § 4 for more details.

## 3. ERROR RATE

We now introduce a level-specific false discovery rate for a general multi-level tree that reflects a hierarchical order of testing, incorporates the logical constraints across the levels in the tree, and preserves the advantages of a marginal error measure. Our error rate assigns a specific role to the error measures within families, adjusting for the data-based selection process that leads to testing those families. It allows us to increase power when the signal is not uniformly distributed across

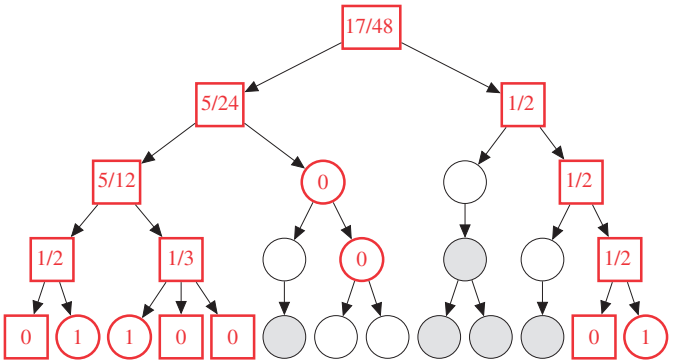


Fig. 2. Illustration of the bottom-up calculation of the proposed error rate for level 4, $\text{sFDR}^4$, using the same configuration of hypotheses as in Fig. 1. The error measure $\mathcal{E}_j(4)$ is defined for rejected hypotheses and indicated by the red number in the node corresponding to each rejection. The hypotheses not distinguished by red borders are not rejected and so do not receive any error measure. If the rejections are nodes at the level of interest, which is level 4 in this illustration, the error measure is 1 for an incorrect rejection and 0 otherwise. For a node at a higher level, the error measure is the average of the error measures assigned to its children if it has one or more rejected child hypotheses and is 0 otherwise.

the tree of hypotheses, and it accounts for the entire testing strategy that leads to the findings at the resolution of interest.

We start by giving a recursive definition of our error rate, as illustrated in Fig. 2. To define an error measure for level $\ell$, we begin by assigning Type I error indicators to the hypotheses rejected at level $\ell$; that is, for each rejected $H_i$, $\mathcal{E}_i(\ell) = I$, where $H_i$ is null. Moving upwards to coarser resolutions, each rejected hypothesis $H_j$ is assigned an error measure $\mathcal{E}_j(\ell)$ that is the average of the error measures assigned to its rejected children or 0 if it has no rejected children. So, for example, at level $\ell - 1$ we define as the error of a parent hypothesis the proportion of false discoveries among its children. Hence, the error assigned to a hypothesis at level $\ell - 1$ is 1 when all discoveries among its children are false, is 0 when none is a false discovery, and can take values in between. Generally, letting $\mathcal{S}_j$ be the set of indices of rejected hypotheses in the family $\mathcal{F}_j$, where we have suppressed the superscript indicating the level of the family, we define

$$\mathcal{E}_j(\ell) = \frac{\sum_{r \in \mathcal{S}_j} \mathcal{E}_r(\ell)}{|\mathcal{S}_j|},$$

where the sum over an empty set is evaluated as 0 and the fraction $0/0$ is defined to be 0. The proposed selective false discovery rate $\text{sFDR}^\ell$ is obtained by iterating this process until the root hypothesis $\mathcal{E}_0(\ell)$ is reached and then taking the expectation of its error measure:

$$\text{sFDR}^\ell = E\{\mathcal{E}_0(\ell)\}.$$

If the root node is not rejected, no hypothesis is rejected, so in this case $\mathcal{E}_0(\ell)$ is trivially equal to 0. Writing explicitly the expression for $\mathcal{E}_0(\ell)$, denoted by $\text{sFDP}^\ell$ hereafter, we can see how it is the weighted average of false discovery proportions in tested families at level $\ell$, with the weights depending on the numbers of rejected hypotheses along the path leading to the families:

$$\text{sFDP}^\ell = \sum_{i \in \mathcal{S}^{\ell-1}} w_i^\ell \, \text{FDP}(\mathcal{F}_i^\ell)$$

for $\ell \geqslant 1$. Here $\mathcal{S}^{\ell-1}$ is the set of indices of the rejected hypotheses at level $\ell - 1$, $\text{FDP}(\mathcal{F}_i^\ell)$ is the false discovery proportion within $\mathcal{F}_i^\ell$, defined as 0 if no hypothesis within $\mathcal{F}_i^\ell$ is rejected, and

$$w_i^\ell = \frac{1}{\prod_{j \in A(i)} |\mathcal{S}_j|} \tag{2}$$

for $\ell > 1$ while $w_0^1 = 1$, with $A(i)$ denoting the set of indices of ancestors of $H_i$. In other words, $w_i^\ell$ for $\ell > 1$ is the inverse of the product of the numbers of rejections in the families that include ancestors of the family $\mathcal{F}_i^\ell$. This definition makes explicit how the random weights for the false discovery proportion of the family $\mathcal{F}_i^\ell$ are adaptive to the number of rejections along the branch that resulted in the selection of the family. It also allows us to see the similarity between sFDR and more traditional false discovery rates. The weight of the level-1 family $\mathcal{F}^1$ is always 1, and so if the root node hypothesis is always rejected, then $\text{sFDR}^1$ is equal to $\text{FDR}^1$, the level-1 restricted false discovery rate. Similarly, $\text{sFDR}^2$ is equal to the error criterion introduced in Benjamini & Bogomolov (2014), i.e., the expected average false discovery proportion across all the selected level-2 families.

Alternative weighting schemes, such as equal weights, could also be used. We choose to focus on adaptive weights, however, as they account for heterogeneity in the amount of signal across different branches in the tree: as the number of rejections in the path leading to family $\mathcal{F}_i^\ell$ increases, the selection effect is less pronounced, and the cost of a false discovery in $\mathcal{F}_i^\ell$ decreases. Like the false discovery rate, which is adaptive to the amount of signal within a family, becoming less stringent when there are many effects, $\text{sFDR}^\ell$ is adaptive to the amount of signal in the selected families at level $\ell$ and their ancestor families.

In the Supplementary Material we illustrate the computation of $\text{sFDP}^\ell$ as the weighted random average in (2) for the tree in Figs. 1 and 2. The signals are more concentrated in level-2 and level-3 families that are descendants of $H_1$ than in families that are descendants of $H_2$, which leads to more rejections within the former families. This results in smaller weights for the false discovery proportions of selected level-4 families whose ancestor is $H_1$. For example, the weight for the false discovery proportion within the family $\{H_{14}, H_{15}\}$ is $1/8$, and the weight for $\{H_{25}, H_{26}\}$ is $1/2$, while the false discovery proportion is the same for both families. The selection process leading to the latter family is more stringent, leading to a higher price for the false discovery proportion within that family. The global false discovery proportion for level-4 discoveries is $3/7$; however, it does not account for the fractions of errors within the selected families. See the Supplementary Material for additional examples illustrating the motivation for controlling $\text{sFDR}^\ell$ rather than the global level-$\ell$ false discovery rate.

*Remark* 1. The definition of $\text{sFDR}^\ell$ can be extended to error measures other than the false discovery proportion within the selected families. Specifically, for each selected family $\mathcal{F}_i^\ell$, one can replace $\text{FDP}(\mathcal{F}_i^\ell)$ in (2) by $\mathcal{C}_i(\mathcal{F}_i^\ell)$, where $\mathcal{C}_i$ is a discrepancy measure such that $E(\mathcal{C}_i)$ is a known error rate, e.g., the weighted false discovery rate, familywise error rate or false discovery exceedance $\text{pr}\{\text{FDP}(\mathcal{F}_i^\ell) > \gamma\}$. Moreover, $\text{sFDR}^\ell$ admits any hierarchical selection process leading to selection of families at level $\ell$. For example, the selection rule may reduce to selecting each hypothesis at level $k \leqslant \ell - 1$ if the minimum $p$-value within the family it indexes is below 0.05, and its parent hypothesis is selected, regarding the root hypothesis as being always selected. Although any hierarchical selection process is admitted, the concept of $\text{sFDR}^\ell$, which gives higher penalties for false discoveries within families that are more carefully selected, makes sense only for the selection rules that choose the most promising families at level $\ell$. Such selection rules appear quite reasonable in practice.

## 4. Testing strategy

### 4.1. *The TreeBH procedure*

TreeBH is a hierarchical testing strategy that targets the control of $\text{sFDR}^{\ell}$ to a prespecified bound $q^{(\ell)}$ for all $\ell \in \{0, \ldots, L\}$ simultaneously. The name reflects the fact that it extends the Benjamini–Hochberg procedure to the context of tree-structured hypotheses. Testing is conducted in a stepwise fashion, starting from the root node. At each level of the tree, the Benjamini–Hochberg procedure is applied separately to each selected family of hypotheses, with a working target that is more stringent if fewer rejections were made in previous steps. If there is a hypothesis of interest corresponding to the root node, the procedure starts by testing the root node at level $q^{(0)}$, and proceeds to the following steps only if this hypothesis is rejected. If there is no such hypothesis, the root node is considered to be always rejected, and the procedure starts from Step 1. The procedure is defined as follows.

*Step* 1. Apply the Benjamini–Hochberg procedure with the bound $q^{(1)}$ to the $p$-values of family $\mathcal{F}_0^1$. If $\mathcal{S}_0^1 = \phi$, i.e., if no hypothesis is rejected in $\mathcal{F}_0^1$, then stop; otherwise, proceed to Step 2.

*Step* 2. For $\ell \in \{2, \ldots, L\}$, apply the following steps sequentially.

*Step* $\ell$. For each family $\mathcal{F}_i^{\ell}$ whose parent $H_i$ is rejected, apply the Benjamini–Hochberg procedure to the $p$-values of family $\mathcal{F}_i^{\ell}$ with working target $q^{(\ell)}$ multiplied by the product of proportions of rejected hypotheses within the families that contain ancestors of the family $\mathcal{F}_i^{\ell}$:

$$q_i = \left\{ \prod_{j \in A(i)} \frac{|\mathcal{S}_j|}{|\mathcal{F}_j|} \right\} q^{(\ell)} = q_{\mathrm{P}(i)} \frac{|\mathcal{S}_{\mathrm{P}(i)}|}{|\mathcal{F}_{\mathrm{P}(i)}|} \frac{q^{(\ell)}}{q^{(\ell-1)}},$$

where $j = \mathrm{P}(i)$ if $H_j$ is the parent of $H_i$, and $q_0 = q^{(1)}$. If no rejections are made or if $\ell = L$, then stop; otherwise, proceed to step $\ell + 1$.

The working target for level-2 tested families is equal to $q^{(1)}$ multiplied by the proportion of selected level-1 hypotheses, $|\mathcal{S}_0^1|/|\mathcal{F}_0^1|$. This is the adjustment suggested by Benjamini & Bogomolov (2014) for control of $\text{sFDR}^2$, accounting for selective inference on level-2 families, which is the procedure we generalize here. For families at the same level $\ell > 2$, the targeted bound may be different, depending on the proportion of selected hypotheses within their ancestor families. For example, for the rejection pattern illustrated for the tree in Fig. 1, the family $\{H_{14}, H_{15}\}$ is tested using the Benjamini–Hochberg procedure with bound $q^{(4)}$, because all the hypotheses are rejected within the families of its ancestors, while the family $\{H_{25}, H_{26}\}$ is tested with bound $q^{(4)}/4$, because the proportions of rejections within the families containing ancestors of $\{H_{25}, H_{26}\}$ are 1, 1/2 and 1/2. This corresponds to the value of $\text{sFDP}^{\ell}$.

One can replace the Benjamini–Hochberg procedure by another procedure in each step. For example, the power of TreeBH can be improved by applying different variants of the Benjamini–Hochberg procedure within the selected families, possibly incorporating prior-knowledge weights on the hypotheses (Genovese et al., 2006) or the estimator of the proportion of nulls within a family (Storey et al., 2004; Ramdas et al., 2019). More radically, one can adopt a procedure controlling an error rate other than the false discovery rate, as discussed in Remark 1. The theoretical properties of the TreeBH procedure and its extensions are described in the next subsection.

### 4.2. *Error rate control*

Before examining under what conditions TreeBH controls $\text{sFDR}^{\ell}$, we state a property relating to the consistency of discoveries across levels. The fact that $\text{sFDR}^{\ell}$ incorporates information on the hierarchical order of testing leads to coherence between the findings at different levels: a discovery at a finer scale can happen only if the corresponding coarser hypothesis has been rejected. If we impose an additional consonance condition, requiring that whenever a parent hypothesis is rejected, at least one of the hypotheses in the family it indexes must be rejected, we can obtain the following result that derives control of $\text{sFDR}^{\ell-1}$ from the control of $\text{sFDR}^{\ell}$ at the higher level for a general hierarchical testing procedure.

PROPOSITION 1. *The following properties hold:*

(i) *If a hierarchical testing procedure is consonant, then for each* $\ell \geqslant 1$ *we have that* $\text{sFDR}^{\ell-1} \leqslant \text{sFDR}^{\ell}$, *with equality when the proportion of true null hypotheses in each family at level* $\ell$ *is either* 0 *or* 1.

(ii) *When the p-value for each parent hypothesis is computed using Simes' method applied to the p-values within the family it indexes and the targeted* $\text{sFDR}^{\ell}$ *bounds satisfy* $q^{(0)} \leqslant q^{(1)} \leqslant \cdots \leqslant q^{(L)}$, *it is guaranteed that the TreeBH procedure is consonant.*

We are now ready to state precisely when TreeBH results in control of $\text{sFDR}^{\ell}$. We consider assumptions pertaining to the dependency between the $p$-values, the tree structure, the relations between the hypotheses and between their $p$-values, and error and procedure properties. To study false discovery rate control, the following notion of positive dependence has proven useful.

DEFINITION 1 (Benjamini & Yekutieli, 2001). *The vector* $X = (X_1, \ldots, X_m)$ *is positive regression dependent on a subset* $I_0 \subseteq \{1, \ldots, m\}$ *if for any increasing set* $D$ *such that* $x \in D$ *and* $y \geqslant x$ *coordinatewise imply* $y \in D$, *and for each* $i \in I_0$, $\text{pr}(X \in D \mid X_i = x)$ *is nondecreasing in* $x$.

Consider the following dependency structures.

*Assumption* 1. The vector of $p$-values for the hypotheses at the finest level $L$ is positive regression dependent on a subset of indices corresponding to true null hypotheses.

*Assumption* 2. The $p$-values in each family at level $L$ are independent of the $p$-values in any other family at level $L$.

Consider the following assumptions on the structure and construction of the tree.

*Assumption* 3. The $p$-value for each parent hypothesis $H_i$ is a combination of the $p$-values within the family it indexes, satisfying $P_i = f_i(P_{\mathcal{F}_i})$, where $P_{\mathcal{F}_i}$ is the vector of $p$-values for the family $\mathcal{F}_i$ and $f_i : [0,1]^{|\mathcal{F}_i|} \to [0,1]$ is a coordinatewise nondecreasing combination function.

*Assumption* 4. Each hypothesis $H_i$ at level $\ell \leqslant L-1$ satisfies $H_i = \bigcap_{j \in \mathcal{F}_i^{\ell+1}} H_j$.

Assumptions 3 and 4 are discussed in the Supplementary Material. To obtain results that are valid in the greatest number of cases, we consider different combinations of the assumptions above. We start with a lemma relative to the highest resolution.

LEMMA 1. *Under Assumptions 1 and 3, the TreeBH procedure guarantees* $\text{sFDR}^{\ell} \leqslant q^{(\ell)}$ *for the finest level* $\ell = L$.

The next theorem deals with the case where each parent hypothesis is the intersection of the hypotheses within the family it indexes, and provides a specific choice of the combination function for which the TreeBH procedure controls $\text{sFDR}^\ell$ for each level $\ell \in \{0, \ldots, L\}$, if the targeted $\text{sFDR}^\ell$ bounds are equal.

THEOREM 1. *If Assumptions 1, 4 and 3 with (1) as the combination function hold and if the targeted* $\text{sFDR}^\ell$ *bounds are equal, i.e., satisfy* $q^{(0)} = q^{(1)} = \cdots = q^{(L)} = q$ *for some* $q \in [0, 1]$, *then the TreeBH procedure guarantees* $\text{sFDR}^\ell \leqslant q$ *for each* $\ell \in \{0, \cdots, L\}$.

*Remark* 2. Similarly to the Benjamini–Hochberg procedure for which $\text{FDR} = q$ when all null hypotheses are true, with independent and uniformly distributed $p$-values, the TreeBH procedure guarantees $\text{sFDR}^\ell = q$ for each level $\ell$ under the same conditions, in the setting of Theorem 1.

We now show how Theorem 1 follows directly from Proposition 1 and Lemma 1. Assume that the assumptions of Theorem 1 hold. Then we obtain that $\text{sFDR}^L \leqslant q$ by Lemma 1. In addition, Proposition 1 shows that in this case the TreeBH procedure is consonant, and $\text{sFDR}^\ell \leqslant \text{sFDR}^L$ for each $\ell \in \{0, \ldots, L\}$, which gives the result of Theorem 1. Using the techniques developed in Ramdas et al. (2019), one can prove that in this case all the $p$-values in the tree are valid.

Theorem 1 concerns the basic yet useful version of the TreeBH procedure. Generalizations of the TreeBH procedure are possible, as discussed in § 4.1. Specifically, following the hierarchical process of TreeBH, each selected family $\mathcal{F}_i^\ell$ may be tested by an $E(\mathcal{C}_i)$-controlling procedure with the same working target $q_i$, possibly replacing the false discovery rate-controlling Benjamini–Hochberg procedure. Consider the following assumptions regarding such a generalized TreeBH procedure.

*Assumption* 5. The error rate $E(\mathcal{C}_i)$ is such that $\mathcal{C}_i$ takes values in a countable set.

*Assumption* 6. Each selected family $\mathcal{F}_i^\ell$ is tested by a procedure that can control $E(\mathcal{C}_i)$ at any target level under the dependency structure therein, and the procedures are simple selection rules, i.e., they satisfy the following condition. For each rejected hypothesis $H_j$ in $\mathcal{F}_i^\ell$, fixing all the $p$-values in the family except $P_j$ and changing $P_j$ so that $H_j$ is still rejected will not change the number of rejected hypotheses in $\mathcal{F}_i^\ell$.

THEOREM 2. *Under Assumption 3 and the dependency satisfying Assumption 2, the generalized TreeBH procedure which satisfies both of Assumptions 5 and 6 guarantees that for each* $\ell \in \{1, \ldots, L\}$,

$$E\left\{\sum_{i \in \mathcal{S}^{\ell-1}} w_i^\ell \mathcal{C}_i(\mathcal{F}_i^\ell)\right\} \leqslant q^{(\ell)} \tag{3}$$

*for* $w_i^\ell$ *defined in (2), where* $\mathcal{C}_i(\mathcal{F}_i^\ell)$ *is the error measure* $\mathcal{C}_i$ *within* $\mathcal{F}_i^\ell$.

General conditions that guarantee the equality in (3) are given in the proof of Theorem 2.

*Remark* 3. Assumption 5 is very lenient. For all common error rates of the form $E(\mathcal{C}_i)$, including the familywise error rate, the false discovery rate and its weighted variant, $\mathcal{C}_i$ takes values in a countable set.

*Remark* 4. Many multiple testing procedures are simple selection rules, as required by Assumption 6. Specifically, any step-up or step-down multiple testing procedure with prespeci-

fied thresholds defines a simple selection rule (Benjamini & Bogomolov, 2014). In particular, the Bonferroni and Benjamini–Hochberg procedures define simple selection rules. Moreover, the adaptive Bonferroni and Benjamini–Hochberg procedures, based on Storey's plug-in estimator for the proportion of true nulls, are also simple selection rules. These procedures have been shown to control the familywise error rate (Finner & Gontscharuk, 2009) and the false discovery rate (Storey et al., 2004) under independence. See Ramdas et al. (2019) for a further extension of the adaptive Benjamini–Hochberg procedure, incorporating weights on the hypotheses, which also defines a simple selection rule and controls the false discovery rate under independence.

Now consider a tree satisfying Assumptions 3 and 4 and also the following assumption.

*Assumption* 7. The $p$-value for each hypothesis at level $L-1$ is computed by Simes' method (1), and the $p$-value for each hypothesis at level $\ell < L-1$ is computed using any combination method that gives a valid global null $p$-value under independence, such as Simes', Fisher's and Stouffer's (Stouffer et al., 1949) methods.

COROLLARY 1. *In the generalized TreeBH procedure for a tree satisfying Assumptions 3, 4 and 7, test each selected level-L family using the Benjamini–Hochberg procedure or its weighted variant suggested by Genovese et al. (2006), and test any other selected family using any multiple testing procedure that guarantees false discovery rate control under independence. When Assumptions 1 and 2 both hold, this generalized TreeBH procedure guarantees that for each* $\ell \in \{1, \ldots, L\}$, $\mathrm{sFDR}^{\ell} \leqslant q^{(\ell)}$.

The dependency structure in Corollary 1 is more restrictive than that in Theorem 1. It therefore allows the use of potentially more powerful methods for combining the $p$-values and for testing the selected families. For general dependence, a more conservative variant of the TreeBH procedure gives the same theoretical guarantees. However, based on our simulation results and the robustness of the Benjamini–Hochberg procedure, we believe that modification of the TreeBH procedure is not required for many types of dependencies encountered in applications.

## 5. EXAMPLES AND SIMULATION

In this section we use an example to illustrate the differences between the error rates and procedures proposed here and other methods in the literature. More simulations can be found in the Supplementary Material. For simplicity we limit ourselves to trees with $L = 3$, where all hypotheses within a given tree level have the same number of children. In these three-level trees we can use an index system for the hypotheses that explicitly indicates logical relations: hypotheses at level 3 are $H_{ijt}$, those at level 2 are $H_{ij\bullet} = \bigcap_t H_{ijt}$, and those at level 1 are $H_{i\bullet\bullet} = \bigcap_j H_{ij\bullet}$. We can then describe the configurations of true and false nulls using a matrix containing all the level-3 hypotheses; see the Supplementary Material. Each row includes all the hypotheses $H_{ijt}$ corresponding to one value of $i$, so that the presence of a nonnull hypothesis in row $i$ signifies that $H_{i\bullet\bullet}$ is false. Each column corresponds to one pair $(j, t)$, with all the columns that have the same value of $j$ being adjacent, so that within a row, blocks of columns correspond to the families in level 3, and the presence of a nonnull hypothesis in row $i$ block $j$ signifies that $H_{ij\bullet}$ is false.

We compare the performance of different approaches in terms of level-specific error rates and power. Specifically, we calculate the false discovery rate for discoveries at levels 1, 2 and 3, denoted by $\mathrm{FDR}^{\ell}$ for $\ell = 1, 2, 3$, as well as the selective $\mathrm{sFDR}^{\ell}$ for levels $\ell = 2, 3$. We omit $\mathrm{sFDR}^1$ since in this example $\mathrm{FDR}^1 = \mathrm{sFDR}^1$. For the hypotheses at each level, we also calculate the power.

Table 1. *Hypothesis configuration for the example, with the nonnull hypotheses marked in red*

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_{1,1,1}$ | $H_{1,1,2}$ | $H_{1,2,1}$ | $H_{1,2,2}$ | $H_{1,3,1}$ | $H_{1,3,2}$ | $H_{1,4,1}$ | $H_{1,4,2}$ | $H_{1,5,1}$ | $H_{1,5,2}$ | $H_{1,6,1}$ | $H_{1,6,2}$ | ... | $H_{1,6,90}$ |
| $H_{2,1,1}$ | $H_{2,1,2}$ | $H_{2,2,1}$ | $H_{2,2,2}$ | $H_{2,3,1}$ | $H_{2,3,2}$ | $H_{2,4,1}$ | $H_{2,4,2}$ | $H_{2,5,1}$ | $H_{2,5,2}$ | $H_{2,6,1}$ | $H_{2,6,2}$ | ... | $H_{2,6,90}$ |
| $H_{3,1,1}$ | $H_{3,1,2}$ | $H_{3,2,1}$ | $H_{3,2,2}$ | $H_{3,3,1}$ | $H_{3,3,2}$ | $H_{3,4,1}$ | $H_{3,4,2}$ | $H_{3,5,1}$ | $H_{3,5,2}$ | $H_{3,6,1}$ | $H_{3,6,2}$ | ... | $H_{3,6,90}$ |
| $H_{4,1,1}$ | $H_{4,1,2}$ | $H_{4,2,1}$ | $H_{4,2,2}$ | $H_{4,3,1}$ | $H_{4,3,2}$ | $H_{4,4,1}$ | $H_{4,4,2}$ | $H_{4,5,1}$ | $H_{4,5,2}$ | $H_{4,6,1}$ | $H_{4,6,2}$ | ... | $H_{4,6,90}$ |
| $H_{5,1,1}$ | $H_{5,1,2}$ | $H_{5,2,1}$ | $H_{5,2,2}$ | $H_{5,3,1}$ | $H_{5,3,2}$ | $H_{5,4,1}$ | $H_{5,4,2}$ | $H_{5,5,1}$ | $H_{5,5,2}$ | $H_{5,6,1}$ | $H_{5,6,2}$ | ... | $H_{5,6,90}$ |
| $H_{6,1,1}$ | $H_{6,1,2}$ | $H_{6,2,1}$ | $H_{6,2,2}$ | $H_{6,3,1}$ | $H_{6,3,2}$ | $H_{6,4,1}$ | $H_{6,4,2}$ | $H_{6,5,1}$ | $H_{6,5,2}$ | $H_{6,6,1}$ | $H_{6,6,2}$ | ... | $H_{6,6,90}$ |

The following methods are included in our comparison: (i) the Benjamini–Hochberg procedure applied across the pooled set of $p$-values for the entire matrix of hypotheses, which guarantees control of $\text{FDR}^3$; (ii) the Benjamini–Bogomolov method applied with hypotheses grouped into a two-level hierarchy with $H_{i\bullet\bullet}$ at level 1, each indexing a family $\mathcal{F}_i^2 = \{H_{ijt} : j = 1, \ldots, m; t = 1, \ldots, k_i\}$, where the selection in level 1 is done by using the Benjamini–Hochberg procedure on Simes' $p$-values for $H_{i\bullet\bullet}$, guaranteeing control of $\text{FDR}^1$; (iii) the $p$-filter applied to the matrix of hypotheses, with groups defined by the pooled set of all hypotheses, rows and columns, which guarantees control of $\text{FDR}^1$; (iv) a hierarchical variant of the $p$-filter applied with groups defined by the pooled set of all hypotheses, rows and sets of columns, in nested fashion, so as to mimic our hierarchical procedure; this guarantees control of $\text{FDR}^\ell$ for $\ell = 1, 2, 3$; (v) TreeBH using three levels, guaranteeing control of $\text{FDR}^1$, $\text{sFDR}^2$ and $\text{sFDR}^3$.

We consider six hypotheses at level 1, each indexing a family of six hypotheses, parents of families at level 3 that contain 2, 2, 2, 2, 2 and 90 hypotheses, respectively, with truth assignments as described in Table 1.

First we discuss the implications of the configuration in Table 1 for error rates. Here five out of six level-1 hypotheses are false, so we can expect $\text{FDR}^1$ to be contained for any method. Consider the error control for level-3 discoveries; methods such as the Benjamini–Hochberg procedure and the $p$-filter, which do not take account of families, will weigh any false discovery against the many possible true discoveries in family $\mathcal{F}_{1,6}^3$, the family that contains 90 nonnull hypotheses. In contrast, for the selective methods we propose, any false discovery in families $\mathcal{F}_{i,6}^3$ for $i = 3, \ldots, 6$ would result in $\text{FDP}_{i,6}^3 = 1$, and this would contribute a substantial weight to the average in $\text{sFDR}^3$.

We next examine how the power of the different procedures is influenced by the configuration in Table 1. A large number of the level-3 families are homogeneous; this gives an advantage to testing procedures that recognize such families, allowing the Benjamini–Hochberg threshold for significance to adapt to the different proportions of nonnull hypotheses. Figure 3 displays the results of a simulation in which all the error rates are targeted using a bound of 0.1, and for each realization the $p$-values $P_i$ for each of the hypotheses are generated independently as follows:

$$X_i \sim \mu + N(0, 1), \quad P_i = 1 - \Phi(X_i),$$

where $\Phi$ denotes the standard normal cumulative distribution function, $\mu = 0$ for null hypotheses, and $\mu > 0$ for nonnull hypotheses, with larger values of $\mu$ corresponding to greater signal strength.

Figure 3 underscores how each of the methods controls its target error rates, but not others. The Benjamini–Hochberg procedure does not control any level-1 or level-2 error rates, nor does it control $\text{sFDR}^3$, and the $p$-filter methods do not control all of the $\text{sFDR}^\ell$. In this set-up it appears that the Benjamini–Bogomolov procedure controls the $\text{sFDR}^\ell$, but we will see in other examples that this is not always the case. Figure 3 also shows how the TreeBH procedure, despite exhibiting the most stringent error control in this example, has the highest power across levels, beaten substantially only by the Benjamini–Hochberg procedure in level 1, where this procedure has no

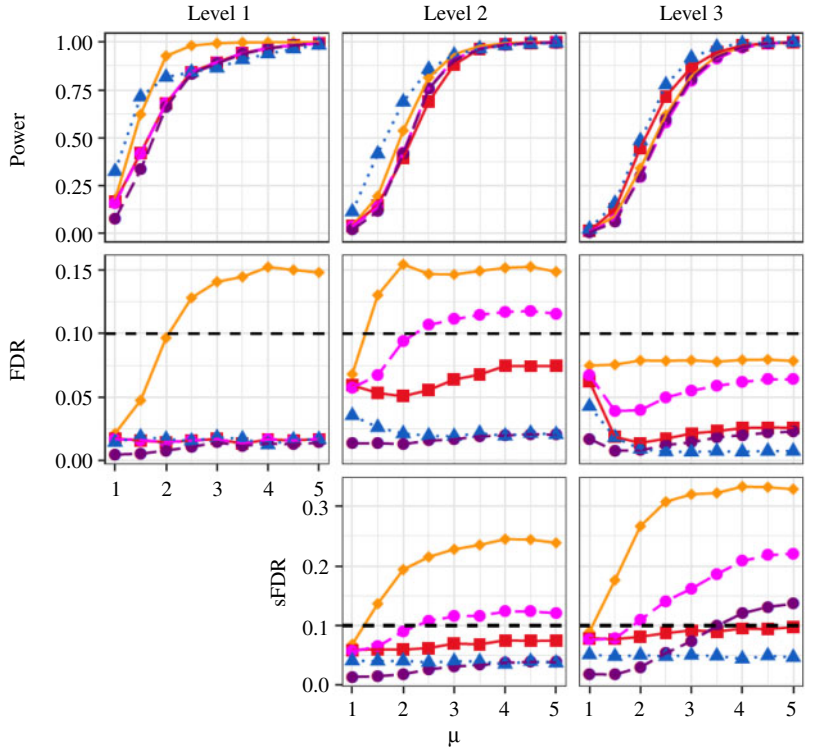


Fig. 3. Results for the example. Each point corresponds to the average of 1000 realizations. Dashed horizontal lines indicate the target values for the error rates. The methods under comparison are the Benjamini–Hochberg procedure (orange diamonds), the Benjamini–Bogomolov method (red squares), the nonhierarchical version of the $p$-filter (pink circles), the hierarchical version of the $p$-filter (purple circles) and TreeBH (blue triangles).

false discovery rate control. Interestingly, its power in levels 2 and 3 is higher than that of the Benjamini–Bogomolov method, which shares some of its hierarchical features and happens to control the selective error rates in this case. The increased power at higher levels is due to the fact that testing is carried out in more homogeneous families. Higher power at level 1 is due to the fact that in calculating the Simes' $p$-values for $H_{i\bullet\bullet}$ one uses six $p$-values rather than 100, and, at least for $i = 3, 4, 5$, five of those are going to be very small, owing to the nonnull status of the hypotheses they represent; the effect of these $p$-values would be more washed out in the entire pool of 100 hypotheses.

The Supplementary Material includes two additional simulation studies: one that facilitates comparison with Foygel Barber & Ramdas (2016), and one that illustrates the applicability of our procedure to multi-trait genetic association studies. In the first of these simulations, the nonhierarchical $p$-filter, although controlling the level-1 false discovery rate, fails to control the level-2 false discovery rate or the level-2 and level-3 selective error rates. The hierarchical version of the $p$-filter is able to control all error rates considered, but achieves lower level-3 power than does TreeBH. The TreeBH method, on the other hand, yields the best overall performance in terms of error control and power, even if many of the differences are small. The second simulation in the Supplementary Material studies the association of more than 8000 genetic variants with the expression of 250 genes in each of five tissues. By using actual genetic data we obtain a realistic dependence structure among the $p$-values. As the $p$-filter method was not able to finish running in less than 24 hours given similar groupings to those in the simulation with independent $p$-values, we

focused on comparing the Benjamini–Hochberg procedure, the Benjamini–Bogomolov method and TreeBH. Our results show that TreeBH appears to control the targeted error rates, unlike the Benjamini–Hochberg procedure, and it has a similar false discovery rate, sFDR and power to the Benjamini–Bogomolov method.

## 6. Case studies

The first application we consider is the study of genetic regulation across multiple tissues in the human body. The goal of expression quantitative trait loci analysis is to identify DNA variants that influence the expression of genes. Since gene expression levels differ across tissues, expression quantitative trait loci analysis may reveal both shared and tissue-specific patterns of regulation, with important implications for the understanding of disease mechanisms.

Typically, for each tissue $t$, the hypotheses $H_{ijt}$ regarding the association of each gene with nearby genetic variants are tested using a linear model with normalized expression for gene $j$ as the response and the estimated number of copies of the minor allele for variant $i$ as the predictor, and covariates are included to account for potential confounding factors. The most common approach has been to perform error control in each tissue separately (Nica et al., 2011; Grundberg et al., 2012). Results for different tissues are then compared and conclusions are drawn on the tissue-specific nature of the detected associations. However, this approach is prone to error, and joint analysis of multiple tissues is likely to result in lower numbers of false positives and false negatives. Methodology based on meta-analysis (Sul et al., 2013) and Bayesian model selection (Flutre et al., 2013; Li et al., 2017) has been developed to address this shortcoming. The testing procedure we propose here provides some of the advantages of these methods while maintaining the computational benefits of the simpler approach.

One problem of interest in multi-tissue expression quantitative trait loci analysis is to find a set of eSNPs, i.e., single nucleotide polymorphisms, that play a functional regulatory role in at least one tissue. With this goal in mind, we could naturally group the hypotheses into a hierarchical structure with SNPs at level 1, genes at level 2, and tissues at level 3. The level-1 hypothesis $H_{i\bullet\bullet}$ addresses the question of whether SNP $i$ has an effect on expression in any tissue. We consider SNP $i$ to be an eSNP if we reject $H_{i\bullet\bullet}$, and we consider a SNP-gene pair to be discovered if we reject $H_{ij\bullet}$. The $p$-values are defined starting from the leaf hypotheses, which receive the $p$-value from the linear association test. The $p$-values for the level-2 and level-1 hypotheses are then defined using Simes' method. Given this organization of the hypotheses, the TreeBH procedure controls the false discovery rate for eSNPs, the expected average proportion of false SNP-gene associations across the selected SNPs, and the expected weighted average of the proportion of false tissue discoveries for the selected SNP-gene pairs.

Table 2 reports the results of the analysis of a multi-tissue gene expression dataset using the proposed method and a pair of benchmark comparisons. Details of the data, $p$-value computation, procedures, and additional comparisons can be found in the Supplementary Material. The TreeBH procedure is much more conservative at the SNP level, because it provides control of the eSNP false discovery rate, but is less stringent in selecting the genes and tissues for these eSNPs, resulting in a similar number of genes associated with each eSNP to the Benjamini–Hochberg approaches, an increased number of selected tissues for each SNP-gene pair discovered, and a lower percentage of SNP-gene pairs that were discovered in only one tissue. Given the conjecture that local regulatory relationships are likely to be shared across tissues, the TreeBH results seem more biologically plausible than those obtained from the Benjamini–Hochberg methods.

The second application is a microbiome study, concerning the association between gut microorganisms and colorectal cancer. A full description of the data and analysis is given in the Supplementary Material. Here we simply report the results in Fig. 4. This is a clear

Table 2. *Numerical comparison of selection results obtained using the Benjamini–Hochberg procedure applied separately by tissue, the Benjamini–Hochberg procedure applied to the pooled set of p-values from all tissues, and the TreeBH procedure*

| | | Separate | Pooled | TreeBH |
|---|---|---|---|---|
| Level 1 | # eSNPs | 9.1e4 | 8.6e4 | 4.5e4 |
| | % eSNPs | 30% | 28% | 15% |
| Level 2 | # SNP-gene pairs | 1.9e5 | 1.8e5 | 9.3e4 |
| | # genes per eSNP | 2.1 | 2.0 | 2.1 |
| Level 3 | # SNP-gene-tissue triplets | 6.4e5 | 6.2e5 | 5.1e6 |
| | # tissues per SNP-gene pair | 3.3 | 3.5 | 5.4 |
| | % SNP-gene pairs 1 tissue only | 61% | 61% | 48% |

# eSNPs, number of selected SNPs; % eSNPs, selected SNPs as a percentage of the total number of SNPs tested for association; # SNP-gene pairs, number of associated SNP-gene pairs discovered; # genes per eSNP, average number of genes per eSNP across all discovered eSNPs; # SNP-gene-tissue triplets, number of associated SNP-gene-tissue triplets discovered; # tissues per SNP-gene pair, average number of tissues per SNP-gene pair across all discovered SNP-gene pairs; % SNP-gene pairs 1 tissue only, percentage of associated SNP-gene pairs that were discovered in only one tissue.

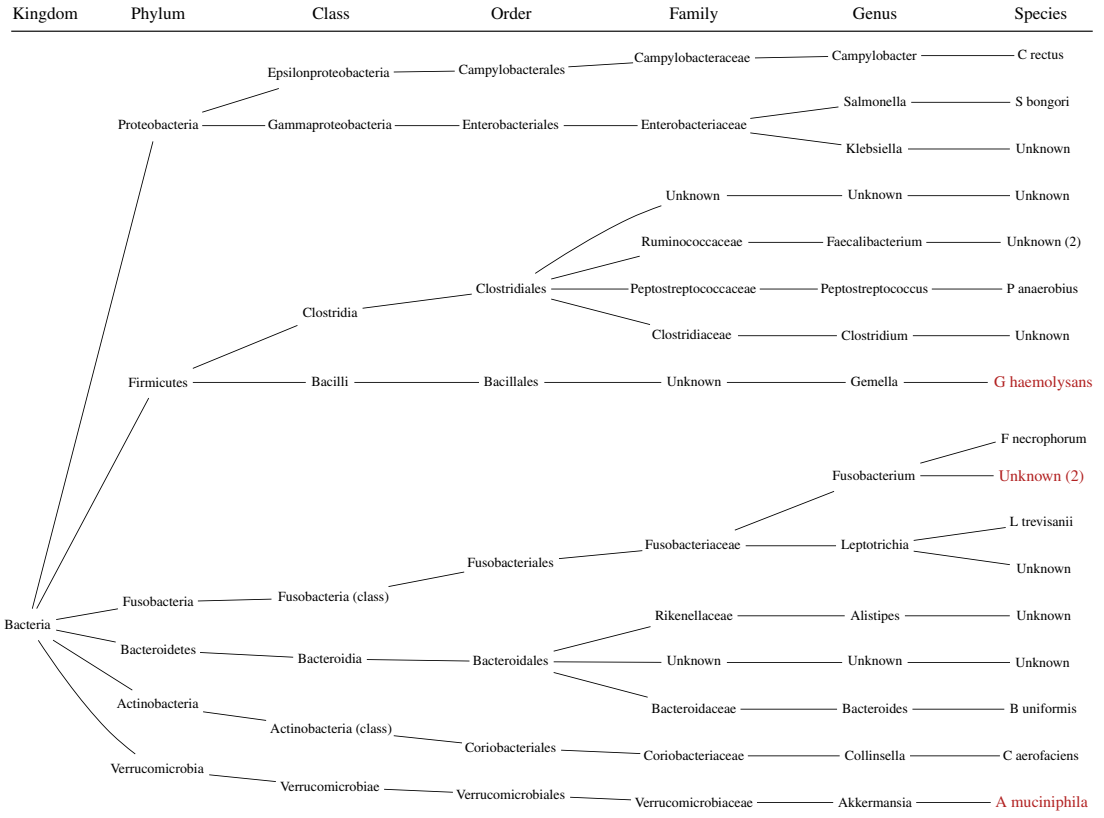


Fig. 4. Taxonomic tree of selections obtained using the TreeBH procedure. Additional discoveries of TreeBH that were not found with the Benjamini–Hochberg procedure are marked in red.

example of a collection of hypotheses that can be organized on a tree, where scientists are interested in responses at multiple resolutions corresponding to the levels of the taxonomic tree.

Compared with the Benjamini–Hochberg procedure, TreeBH yields additional discoveries that appear scientifically meaningful, while reducing the number of findings that are difficult to interpret.

## Discussion

The TreeBH procedure introduced here has many aspects in common with the $p$-filter (Foygel Barber & Ramdas, 2016; Ramdas et al., 2019): both procedures control some form of level-specific false discovery rate, and in both of them one assumes that $p$-values for the finer-scale hypotheses are available, which are summarized using a certain combination method to obtain $p$-values for the group-level hypotheses. When adapted to our hierarchical organization of hypotheses, the $p$-filter controls the level-restricted false discovery rate, but takes no account of the distinct families that make up the collection of hypotheses at a given level; therefore, it does not control our selective error rates and cannot gain power by adapting to the different sparsity levels across families. Finally, the computational time required to run the $p$-filter is substantially greater than that for TreeBH, making its application to genomic problems difficult.

Our procedures are extensions of the proposals in Benjamini & Bogomolov (2014) and so have some of the same merits and limitations. In particular, recent work of Heller et al. (2017) has underscored how selective error rates could be controlled with higher power when a conditional approach to testing is possible; the authors demonstrated the feasibility for a two-layer structure and under independence at the second level. In principle, as long as exact conditional distributions can be evaluated, it may be possible to adopt the conditional testing approach of Heller et al. (2017) to also control the selective error rates that we introduced here.

## Supplementary material

Supplementary Material available at *Biometrika* online includes proofs of the theoretical results, the methodology for general dependence, an additional illustration of the proposed error rate, the relation to previous work for two-level trees, further details of the $p$-filter example, the simulation for multi-trait genetic association studies, and the two case studies. Moreover, we provide an implementation of the TreeBH procedure in the `R` (R Development Core Team, 2021) package `TreeBH`.

## Acknowledgement

The authors thank the donors that made possible the GTEx data collection, and all the scientists that participated in the consortium. Benjamini, Petersen and Sabatti acknowledge support from the U.S. National Institutes of Health. Benjamini was also supported by the European Research Council and Bogomolov by the Israel Science Foundation. The authors acknowledge use of the dataset STAMPEED: Northern Finland Birth Cohort 1966 (phs000276.v2.p1). Bogomolov and Peterson contributed equally to this work.

## References

Benjamini, Y. & Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *J. R. Statist. Soc.* B **76**, 297–318.

Benjamini, Y. & Heller, R. (2007). False discovery rates for spatial signals. *J. Am. Statist. Assoc.* **102**, 1272–81.

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

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.

BRZYSKI, D., PETERSON, C. B., SOBCZYK, P., CANDES, E. J., BOGDAN, M. & SABATTI, C. (2017). Controlling the rate of GWAS false discoveries. *Genetics* **205**, 61–75.

FINNER, H. & GONTSCHARUK, V. (2009). Controlling the familywise error rate plug-in estimator for the propotion of true null hypotheses. *J. R. Statist. Soc. B* **71**, 1031–48.

FLUTRE, T., WEN, X., PRITCHARD, J. & STEPHENS, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9**, e1003486.

FOYGEL BARBER, R. & RAMDAS, A. (2016). The *p*-filter: Multi-layer false discovery rate control for grouped hypotheses. *J. R. Statist. Soc.* B **79**, 1247–68.

GENOVESE, C. R., ROEDER, K. & WASSERMAN, L. (2006). False discovery control with *p*-value weighting. *Biometrika* **93**, 509–24.

GOEMAN, J. J. & MANSMANN, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* **24**, 537–44.

GRUNDBERG, E., SMALL, K. S., HEDMAN, Å. K., NICA, A. C., BUIL, A., KEILDSON, S., BELL, J. T., YANG, T-P., MEDURI, E., BARRETT, A. et al. (2012). Mapping cis- and trans- regulatory effects across multiple tissues in twins. *Nature Genet.* **44**, 1084–9.

HELLER, R., CHATTERJEE, N., KRIEGER, A. & SHI, J. (2017). Post-selection inference following aggregate level hypothesis testing in large scale genomic data. *J. Am. Statist. Assoc.* **113**, 1770–83.

HELLER, R., MEIR, A. & CHATTERJEE, N. (2019). Post-selection estimation and testing following aggregate association tests. *J. R. Statist. Soc.* B **81**, 547–73.

LEI, L., RAMDAS, A. & FITHIAN, W. (2020). A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, doi: https://doi.org/10.1093/biomet/asaa064.

LI, G., SHABALIN, A. A., RUSYN, I., WRIGHT, F. A. & NOBEL, A. B. (2017). An empirical Bayes approach for multiple tissue eQTL analysis. *arXiv:* 1311.2948v5.

LYNCH, G. & GUO, W. (2016). On procedures controlling the FDR for testing hierarchically ordered hypotheses. *arXiv:* 1612.04467.

MEINSHAUSEN, N. (2008). Hierarchical testing of variable importance. *Biometrika* **95**, 265–78.

NICA, A. C., PARTS, L., GLASS, D., NISBET, J., BARRETT, A., SEKOWSKA, M., TRAVERS, M., POTTER, S., GRUNDBERG, E., SMALL, K. et al. (2011). The architecture of gene regulatory variation across multiple human tissues: The MuTHER study. *PLoS Genet.* **7**, e1002003.

PERONE PACIFICO, M., GENOVESE, C., VERDINELLI, I. & WASSERMAN, L. (2004). False discovery control for random fields. *J. Am. Statist. Assoc.* **99**, 1002–14.

R DEVELOPMENT CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

RAMDAS, A., CHEN, J., WAINWRIGHT, M. J. & JORDAN, M. I. (2019). A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika* **106**, 69–86.

RAMDAS, A., FOYGEL BARBER, R., WAINWRIGHT, M. J. & JORDAN, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *Ann. Statist.* **47**, 2790–821.

ROSENBAUM, P. R. (2008). Testing hypotheses in order. *Biometrika* **95**, 248–52.

SIEGMUND, D. O., YAKIR, B. & ZHANG, N. (2011). The false discovery rate for scan statistics. *Biometrika* **98**, 979–85.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–4.

STOREY, J. D., TAYLOR, J. E. & 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.

STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. & WILLIAMS JR, R. M. (1949). *The American Soldier: Adjustment During Army Life*, vol. 1 of *Studies in Social Psychology in World War II*. Princeton, New Jersey: Princeton University Press.

SUL, J. H., HAN, B., YE, C., CHOI, T. & ESKIN, E. (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.* **9**, e1003491.

YEKUTIELI, D. (2008). Hierarchical false discovery rate-controlling methodology. *J. Am. Statist. Assoc.* **103**, 309–16.

YEKUTIELI, D., REINER-BENAIM, A., ELMER, G. I., KAFKAFI, N., LETWIN, N. E., LEE, N. H. & BENJAMINI, Y. (2006). Approaches to multiplicity issues in complex research in microarray analysis. *Statist. Neer.* **60**, 414–37.

[*Received on* 26 *October* 2018. *Editorial decision on* 6 *August* 2020]