Testing hypotheses on a tree: new error rates and controlling strategies

Marina Bogomolov*, Christine B. Peterson*, Yoav Benjamini, Chiara Sabatti

May 19, 2017

* These authors contributed equally to this work.
Abstract

We propose a new multiple testing procedure which addresses the challenge of controlling error rates at multiple levels of resolution. Conceptually, we frame this problem as the selection of hypotheses which are organized hierarchically in a tree structure. We provide a detailed algorithm for the proposed sequential procedure, and prove that it controls relevant error rates given certain assumptions on the dependence among the hypotheses. Through simulation, we demonstrate that the proposed strategy controls these error rates in both simple settings and in settings with dependence similar to that encountered in genome-wide association studies, while offering the potential to gain power over alternative methods. Finally, we conclude with two case studies where we apply the proposed method: firstly, to data collected as part of the Genotype-Tissue Expression (GTEx) project, which aims to characterize the genetic regulation of gene expression across multiple tissues in the human body, and secondly, to data examining the relationship between the gut microbiome and colorectal cancer.

Introduction

Our capacity to acquire data has increased tremendously in the last few decades, presenting new opportunities for science, industry, and policy planning. In many cases, the multitude of measurements on many different variables are analyzed using the statistical framework of hypothesis testing: the relatively small number of rejected null hypotheses, among the thousands or millions tested, indicate “discoveries” of novel, possibly unexpected, patterns and connections.

Paradoxically, the theory of statistical hypothesis testing was developed and reached maturity in a fundamentally different context, which is sometimes described today as “hypothesis-driven research”. Fisher’s notion of hypothesis testing was developed at the same time as Popper’s description of the scientific paradigm: researchers formulate one specific hypothesis, collect data to test it, and “hold on to it” until it is rejected on the basis of new observations.

It is remarkable that this framework has been found useful in our present context, where data is gathered before any hypothesis formulation. The meaningful application of statistical hypothesis testing in our “big data” world has been facilitated in large part by the development of novel error rate measures defined on the entire pool of hypotheses tested, and effectively modeling the “fishing expedition” we embark upon when we analyze data on an exhaustive, rather than theory motivated, collection of variables. In particular, a substantial role in addressing the challenges of multiple hypothesis testing has been played by the False Discovery Rate (FDR) \[ \text{FDR} \], the expected proportion of false
rejections among all the rejections in a study. Controlling the FDR is, for example, the accepted standard in the genomic studies of gene expression variation, genetic regulation, and copy number polymorphisms.

Relying on high-throughput technology, we gather data on a very large number of variables, which translates into a very large number of hypotheses tested. Not only we usually do not have strong a priori theory behind these hypotheses, but often, they are at a level of resolution that is not immediately interpretable. Consider, for example, the study of gene expression regulation across tissues, which motivated our research. In projects such as GTEx (the Genotype Tissue Expression project), each of the tested null hypotheses corresponds to the lack of association between one genetic variant and the expression level of one gene in one tissue: there are approximately 20,000-30,000 measured genes, in each of 44 tissues, and genotypes for around 11 million SNPs. The number of possible hypotheses is formidable and, to reduce the multiple testing burden, some restrictions are typically made so that not every one is tested. Even after these simplifications, the most meaningful discoveries for scientists, and the ones that are discussed in the resulting papers, do not correspond to the rejections of single hypotheses: rather than focusing on a specific SNP that is associated to a specific gene in a specific tissue, researchers are interested, for example, in identifying genes whose expression is subject to genetic regulation (eGenes), or SNPs that affect gene expression of possibly more than one gene, in more than one tissue (eSNPs).

This discrepancy between the hypotheses tested and the reported findings can be problematic. While we have fairly versatile strategies to control the FDR, these are crucially based on the way in which discoveries are counted. Practitioners and statisticians alike have repeatedly pointed out that controlling the FDR for the high resolution hypotheses does not guarantee control of the rate of false “aggregate” discoveries. When the resolution of scientific interest is coarser, different strategies have been outlined to control the FDR of these “global” hypotheses starting from tests on the more specific ones. This has been the case, for example, in applications to MRI imaging [2–4], copy number variant detection [5], and genome wide association studies [6]. In other settings, it is arguable that scientists are interested in discoveries at multiple resolutions and error control guarantees should be provided for all of these. This viewpoint motivated our previous work [7–9], the conditional testing approach of [10], and the method recently introduced by [11, 12] known as the p-filter.

The fact that hypotheses can be meaningfully grouped and discoveries can happen at different resolutions presents not only a challenge for FDR control, but also an opportunity to increase power. Yekutieli [13] described a setting where hypotheses can be arranged along a tree, with resolution increasing with distance from the tree node. If one could carry out testing along the tree, one might avoid having to test high resolution
hypotheses when the corresponding global ones do not give any sign of being false: this could reduce the total number of tests (with clear benefits on the multiplicity budget), as well as allow an adaptively higher threshold for discoveries along the branches that appear to contain more signal. Yekutieli [13] points out that there are multiple possible notions of false discovery rate to control and provides bounds for some of these error rates when the Benjamini-Hochberg controlling strategy is applied consecutively to families along the tree and the p-values for all the hypotheses in the tree are independent. Independence between p-values at different levels is a strong requirement: it means, for example, that we cannot use tests for the higher resolution hypotheses to make decisions on the coarser ones. The proposal in [11] does not require independence of p-values across levels, but it does not use a tree structure to increase power. The contributions in [7] and [10] deal with the question of how to test fine resolution hypotheses when their p-values have been used in selecting “promising” groups. We capitalize here on the results in [7] to introduce a novel framework to test families of hypotheses along a tree, where the term “family” is used to refer to a set of scientifically related hypotheses. In particular, the current proposal goes beyond our previous work, which focused on applying a simpler two-stage procedure to genetic studies of multivariate phenotypes [8] and on developing a computational framework for the application of this method to eQTL studies [9]. Here we provide a more general procedure which can accommodate any hierarchical organization of the hypotheses—with multiple levels of resolution, unlike [10]—and, in contrast to competing methods such as [11], leverages the notion of adaptivity discussed above to enable improved power for branches of the tree that contain more signal. We focus on the specific variant of this procedure where the hypotheses within the families are selected by multiple testing and one is interested in error control for each level of the tree. However, our general procedure is very flexible with regard to several aspects: (1) it can accommodate any number of levels of hierarchy, (2) it can focus on families at any specific level of the tree, and control the errors within the selected families at the given level, (3) it can target different error measures, even (4) varying them across different families.

The rest of the paper is organized as follows. In Section 1, we describe the hierarchical testing problem. Section 2 introduces measures of global error that incorporate the selective strategy of testing and illustrates their relations to other interesting rates in the context of a three-level tree. In Section 3 we present strategies that guarantee the control of these error rates. Section 4 provides the definition of error rates and testing strategy for a tree with an arbitrary number of levels. Section 5 contains a series of simulations contrasting the approach we propose to others with related goals as well as exploring the robustness of the controlling strategy to dependence. Finally, Section 6 presents the results from our procedure on the GTEx data set which motivated our study, as well as
results from a microbiome study where the hypotheses are naturally organized into a multi-layer tree.

1 A tree of families of hypotheses

We are interested in problems where the data exploration can be meaningfully carried out at increasing levels of resolution. In the interest of concreteness, we initially focus on situations where the hypotheses can be organized in three levels of specificity; our results (Section 4 and Appendix) hold for an arbitrary number of levels. In Figure 1, a rectangular box identifies hypotheses that are considered as part of a family (i.e. a set of related hypotheses) and are always tested jointly. Each hypothesis at level $\ell$ is parent to a family of hypotheses at level $\ell + 1$. We have one family $F^1$ at level 1 comprising the $n$ hypotheses $H_{i \bullet \bullet}$, each of which indexes a family $F^2_i$ at level 2, with $F^2_i = \{H_{ij \bullet}, j = 1, \ldots, m_i\}$. Finally, each $H_{ij \bullet}$ is parent to a family $F^3_{ij} = \{H_{ijt}, t = 1, \ldots, k_{ij}\}$.

A motivating example for this structure comes from eQTL studies, where $H_{i \bullet \bullet}$ in Level 1 corresponds to the hypothesis that SNP $i$ is not involved in the regulation of the expression of any gene in any tissue. The Level 2 hypothesis $H_{ij \bullet}$ states that SNP $i$ does not affect the expression of gene $j$ in any tissue, while the Level 3 hypothesis $H_{ijt}$ specifies that SNP $i$ is not involved in the regulation of gene $j$ in tissue $t$. Discoveries in Level 1 correspond to the identification of important SNPs, in Level 2 of the affected genes, and in Level 3 of the tissues where this regulation is active. We will go back to the analysis of eQTL data in the last section of the paper: here we simply want to
emphasize that hierarchical organization of hypotheses is meaningful in that context, with a number of relevant hierarchies that could be defined.

We assume that valid \( p \)-values \( p_{ijt} \) for the Level 3 hypotheses have been computed. P-values for hypotheses at Level 2 and 1 could be available, but in general we consider that they will be obtained from the p-values of the hypotheses in Level 3, thereby not requiring any independence across levels. Unless specified otherwise, we derive \( p \)-values for the Level 2 hypotheses using Simes method \[14\] on the p-values of the family of hypotheses they index at Level 3, and, in turn, use these to obtain \( p \)-values for Level 1 hypotheses again using Simes method:

\[
p_{ij \cdot} = \min_t p_{ij(t)} \times \frac{k_{ij}}{t}
\]
\[
p_{i \cdot \cdot} = \min_j p_{i(j) \cdot} \times \frac{m_i}{f},
\]

where indices in parentheses signify that the p-values have been sorted in increasing order. We rely on the Simes method to compute the \( p \)-value for the parent intersection null hypothesis as it is relatively robust to dependence and has nice properties when used in conjunction with the BH procedure; alternative methods, such as Fisher or permutation, could also be used. We note again that these assumptions mirror well the setting of eQTL studies, where data analysis starts with the computation of the Level 3 p-values \( p_{ijt} \), for which there are a number of efficient algorithms, including Matrix eQTL \[15\] and FastQTL \[16\].

We develop testing procedures that proceed along the tree, from coarser to finer levels. The Level 1 family \( \mathcal{F}^1 \) is always tested; from Level 2 onwards, a family of hypotheses is tested only if its parent hypothesis is rejected. This strategy is similar to that of Yekutieli (2008) \[13\].

2 Global error definitions

Once the hypotheses have been organized in this hierarchical structure, it is clear that discoveries can be made at multiple levels and it becomes necessary to identify a meaningful notion of global error to control. By definition, FDR is a measure relative to the total number of discoveries: depending on what what we consider the relevant total we have a different error rate. Yekutieli \[13\], for example, discusses three choices: the full tree FDR (based on counting all discoveries at each level), the level-restricted FDR (in which one focuses on findings at a specific level), and the outer node FDR (which considers only the finest scale results reached, that is all discoveries that are not parents to other discoveries). While for the single researcher, who has committed to a hierarchical
study of the data, the outer node discoveries (and hence the outer node FDR) might be the most interesting, level specific error rates might be easier to interpret and the results of a procedure that control these might be easier to share broadly with the scientific community, even in the absence of a general consensus on the opportunity of testing hypotheses hierarchically. On the other hand, separately controlling the FDR at each level can be an inadequate way of handling multiplicity (specifically when the hierarchy includes many layers, instead of the three we consider) and, more importantly, may lead to results that lack coherence across resolutions. The p-filter methodology, introduced in [11] and not tied to hierarchical testing, when adapted to our setting, suggests controlling the level-specific FDRs, but after imposing consistency across different levels: coarser discoveries happen if and only if at least one of the finest level hypotheses that they index is rejected.

We introduce here a new notion of level-specific FDR that reflects a hierarchical order of testing where families of hypotheses at higher resolution are tested only when their parent hypotheses have been rejected. This resolves issues of consistency across levels, while preserving the appeal of marginal error measures. Building on the work in [7], our error rate assigns a specific role to the families $\mathcal{F}_s^\ell$ tested at level $\ell$. We suggest controlling the expectation of the weighted average false discovery proportion across the families tested at each level, with the number and identity of families tested random outcomes of the testing that took place at previous levels. Formally,

$$
\text{sFDP}_\ell = \sum_{\mathcal{F}_i^\ell \text{ tested}} w_i^\ell \text{FDP}_{\mathcal{F}_i^\ell}
$$

$$
\text{sFDR}_\ell = \mathbb{E}(\text{sFDP}_\ell)
$$

where $w_i^\ell$ are random weights for FDP$_{\mathcal{F}_i^\ell}$ that sum to 1. The weights can be simply equal to the inverse of the total number of tested families at level $\ell$, so that the error measure is a simple average, i.e.

$$
\text{sFDP}_\ell = \frac{\sum_{\mathcal{F}_i^\ell \text{ tested}} \text{FDP}_{\mathcal{F}_i^\ell}}{|\{i : \mathcal{F}_i^\ell \text{ tested}\}|'}
$$

defined as zero if no family at level $\ell$ is tested, or can be adaptive to the number of rejections along the branch that resulted in the selection of the family $\mathcal{F}_i^\ell$. Considering simple averages for Levels 1 and 2, sFDR$^1$ is equal to FDR$^1$, the Level 1 restricted FDR, as there is only one family at Level 1. Similarly, sFDR$^2$ is equal to the error rate introduced in [7] of $\mathbb{E}(\sum_{i \in S^1} \text{FDP}_i / \max\{|S^1|, 1\})$, with $S^1$ the collection of families $\mathcal{F}_i^2$ indexed by the rejected hypotheses $H_{i\bullet\bullet}$ at Level 1 and FDP$_i$ the false discovery proportion within
\( \mathcal{F}^3_i \). At the third level, we consider two possible weights for the relevant average:

\[
\text{sFDR}^3_T = \mathbb{E} \left( \frac{1}{\max\{|S^1_i|, 1\}} \sum_{i \in S^1} \frac{1}{\max\{|S^2_i|, 1\}} \sum_{j \in S^2_i} \text{FDP}_{ij} \right),
\]

(2)

\[
\text{sFDR}^3_e = \mathbb{E} \left( \frac{1}{\max\{\sum_{k \in S^1} |S^2_k|, 1\}} \sum_{i \in S^1} \sum_{j \in S^2_i} \text{FDP}_{ij} \right),
\]

(3)

where FDP_{ij} is the FDP within \( \mathcal{F}^3_i \), \( S^2_i \) is the set of indices of rejected hypotheses in \( \mathcal{F}^3_i \). \( \text{sFDR}^3_T \) incorporates weights which are adaptive to the process of selection along the tree, while \( \text{sFDR}^3_e \) uses equal weights.

The \( \text{sFDR}^\ell \) are selective error rates in that they measure the average incidence of errors within the selected families, similar to the FDR, which measures the average incidence of type I errors among the selected hypotheses. Moreover, the \( \text{sFDR}^\ell \) incorporate the selective process with which hypotheses come to be considered by the researcher, which may be based on multiple testing, or on other data dependent selection rules. They are marginal error rates, in that the effect of selection is not accounted for by conditioning on the specific results at higher levels. Finally, it is important to note that the \( \text{sFDR}^\ell \)'s assign an important role to families: hypotheses in the same family are considered as addressing the same scientific question, and guarantees on the family-specific FDRs are achieved.

Note that when the numbers of selected hypotheses within all the selected families at Levels 2 are equal, the realized error measures \( \text{sFDP}^3_T \) and \( \text{sFDP}^3_e \) assign equal weights to all the tested families at Level 3, and both error measures are the average FDP across all the tested families at Level 3. In other cases, there may be a big difference between the error measures. If the families across different branches of the tree are very heterogenous in terms of signal, \( \text{sFDR}^3_T \) and \( \text{sFDR}^3_e \) may be very different: \( \text{sFDR}^3_T \) will account less for errors in the families whose ancestor families are enriched with signal, compared to families whose ancestor families contain little signal; \( \text{sFDR}^3_e \), however, treats all the families equally and does not account for the heterogeneity of the selected families at Level 2.

We have emphasized how the fact that the \( \text{sFDR}^\ell \) incorporate in their definition information on the hierarchical order of testing immediately leads to some 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 further impose an additional consistency, requiring that whenever a parent hypothesis is rejected at least one of the hypotheses in the family it indexes is 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.
Proposition 2.1. Consider the two following conditions for a hierarchical testing procedure for the hypotheses in Figure 1:

1. For any \( i \in S^1 \), at least one of the hypotheses in \( F_1^i \) is rejected, i.e. \( S_1^2 \) is not empty.

2. For any \( j \in S_1^2 \) at least one hypothesis in \( F_3^i_j \) is rejected.

If condition 1 holds, then control of \( sFDR_2 \) implies control of \( sFDR_1 \). If condition 1 and 2 hold, then control of \( sFDR_3^T \) implies both control of \( sFDR_2 \) and FDR. A detailed proof is given in the Appendix and partially based on [7]. The result rests on the following relationships between the error rates, which hold when conditions 1 and 2 are satisfied:

\[
sFDR^1 = E \left( \frac{|S_1 \cap \{ i : H^i_{i_0} \text{ is true} \}|}{\max{|S_1|,1}} \right)
\]

\[
sFDR^2 = E \left( \frac{|S_1 \cap \{ i : H^i_{i_0} \text{ is true} \}|}{\max{|S_1|,1}} \right) + E \left( \frac{\sum_{i \in S_1 \cap \{ i : H^i_{i_0} \text{ is false} \}} FDP_i}{\max{|S_1|,1}} \right)
\]

\[
sFDR_3^T = E \left( \frac{|S_1 \cap \{ i : H^i_{i_0} \text{ is true} \}|}{\max{|S_1|,1}} \right) + E \left( \frac{\sum_{i \in S_1 \cap \{ i : H^i_{i_0} \text{ is false} \}} FDP_i}{\max{|S_1|,1}} \right)
\]

\[
+ E \left( \frac{1}{\max{|S_1|,1}} \sum_{j \in S_2^i, H_{ij} \text{ is false}} \sum_{i \in S^1 \cap \{ i : H^i_{i_0} \text{ is false} \}} \frac{1}{\max{|S_2^i|,1}} FDP_{ij} \right)
\]

When the consistency conditions hold, then, the error rate for each level is more strict than the error rates for the previous levels. This implies that in order to control \( sFDR^\ell \) in each of the levels, one may target control of \( sFDR_3^T \) and use any hierarchical selection strategy (i.e. the selection of hypotheses in Levels 1 and 2 does not have to be the result of hypothesis testing).

3 Testing strategy for a three level tree

Let \( q^{(1)} \), \( q^{(2)} \), and \( q^{(3)} \) represent the error rates targeted in each level of the tree.

Testing procedures: 3-level TreeBH and Repeated BB

Step 1. Apply the BH [1] procedure at level \( q^{(1)} \) to \( p_1, \ldots, p_n \) and let \( S_1 \) be the set of rejected \( F^1 \) hypotheses.
Step 2. For each $i \in S^1$, apply the BH procedure at level $|S^1|q(2)/n$ to the family $F^2_i$ p-values $p_{i1}, \ldots, p_{im}$. Let $S^2_i$ be the set of rejected hypotheses in family $F^2_i$.

Step 3. For each $i \in S^1$ and $j \in S^2_i$, apply the BH procedure to $p_{ij1}, \ldots, p_{ijt}$ at adjusted level $q_{ij}^{(3)}$. There are two options for $q_{ij}^{(3)}$. The first one, which leads to the procedure we call TreeBH is $q_{ij}^{(3)} = |S^1||S^2_i|q^{(3)}/(nm_i)$. The second option, which defines the procedure we call Repeated BB (R-BB) is $q_{ie}^{(3)} = \sum_{i \in S^1} |S^2_i|q^{(3)}/\sum_{i=1}^n m_i$.

The choice of $q_{ij}^{(3)}$ in the procedure is such that the penalization depends on the number of rejections in families $F^2_i$ and $F^1$ only, while $q_{ie}^{(3)} = q_e^{(3)}$ is not a function of $i$ as it is based on the total number of families tested at Level 3. The name Repeated BB for the choice of $q_e^{(3)}$ is motivated by the fact that this procedure applies repeatedly the procedure in [7], (which we refer to as BB): the first time at Step 2, considering the errors within the selected families at Level 2 and adjusting for selection at Step 1, and the second time at Step 3, considering the errors within all the selected families at Level 3 and adjusting for the combined selection of Steps 1 and 2. In the eQTL mapping problem we consider in the current work, choosing $q_{ij}^{(3)}$ over $q_{ie}^{(3)}$ would make it easier to discover the effect of a SNP in multiple tissues if the SNP has been identified as influencing multiple genes. Furthermore, the TreeBH procedure has the following consistency property:

Remark 1. When Simes’ p-values are used for constructing p-values for Steps 1 and 2, the TreeBH procedure guarantees that for any $i \in S^1$, $S^2_i \neq \emptyset$ (i.e. for any selected Level 1 hypothesis at least one corresponding Level 2 hypothesis is rejected), and for any $j \in S^2_i$ (i.e. for any selected Level 2 hypothesis), at least one Level 3 discovery is made.

Combining the theory given in the next section, and the results in [7], we obtain the following (see Appendix for the proof).

Theorem 1. If the $p$-values in each family at Level 3 are independent of the $p$-values in any other family at Level 3, and the BH procedure is valid for the dependence structure of the $p$-values within each family at Level 3, then the TreeBH and the Repeated BB procedures guarantee:

1. $\text{FDR}^1 \leq q^{(1)}$
2. $\text{sFDR}^2 \leq q^{(2)}$
3. The choice of $q_{ij}^{(3)}$ in Step 3, leading to TreeBH procedure, assures that $\text{sFDR}^3_T < q^{(3)}$.
4. The choice of $q_{ie}^{(3)}$ in Step 3, leading to Repeated BB procedure, assures that $\text{sFDR}^3_e < q^{(3)}$. 
The 3-level hierarchical testing strategy we describe here is implemented for application in the context of eQTL studies in the R package TreeQTL, available online at http://www.bioinformatics.org/treeqtl/. The R code for the more general multi-layer selection procedure will be made available as well in association with the publication of this manuscript.

4 General trees

We here state our results in the general case of hierarchical testing along a tree with an arbitrary number $L$ of levels. To avoid cumbersome notation, each of the $T$ hypotheses is identified by a simple index $i$ and the information on how they are organized in families on different levels of the tree is stored separately: let $\mathcal{M} = \{H_i, i = 1, \ldots, T\}$ be the complete collection of hypotheses and let $\mathcal{T}$ be a tree that describes how they are arranged in families across $L$ levels. As before, $S^1$ denotes the indices of the selected hypotheses at Level 1, and $S^k_j$ denotes the rejected hypotheses within $\mathcal{F}^k_j$, the family indexed by $H_j$ at level $k − 1$.

4.1 Selective error rates

To obtain a definition $sFDR^\ell$, we generalize $sFDR^3$ as follows

$$sFDP^\ell = \sum_{i_0 \in S^0} \frac{1}{|S^0|} \sum_{i_1 \in S^1} \max\{|S^1|, 1\} \sum_{i_2 \in S^2_{i_1}} \max\{|S^2_{i_1}|, 1\} \cdots \sum_{i_{\ell-1} \in S^{\ell-1}_{i_{\ell-2}}} \max\{|S^{\ell-1}_{i_{\ell-2}}|, 1\} \frac{1}{\mathbf{FDP}^{\mathcal{F}^\ell_{i_{\ell-1}}}}$$

$$sFDR^\ell = \mathbb{E}(sFDP^\ell), \quad (4)$$

where for ease of notation we artificially add a root node hypothesis $H_{i_0}$ which is considered as a parent of $\mathcal{F}^1$ and is always selected, i.e. $S^0 = \{i_0\}$, and $\mathcal{F}^1 = \mathcal{F}^1_{i_0}$.

Remark 1  Note that the definition (4) can be extended to consider error rates other than FDR. Specifically, for each selected family $\mathcal{F}^\ell_i$, one may replace $\mathbf{FDP}^{\mathcal{F}^\ell_i}$ by an error measure $C^\ell_i$ such that $\mathbb{E}(C^\ell_i)$ is a known error rate, e.g. FWER or $\mathbb{P}(FDP > \gamma) = \mathbb{E}(I_{\{FDP > \gamma\}})$ [7]. The procedure and theory for this case is given in the Appendix.

Remark 2  Note that simple FDR for one family of hypotheses is the expected average across all the type I error indicators for the selected hypotheses. Interestingly, $sFDR^\ell$ can be viewed as a generalization of the simple FDR for a tree of hypotheses, when we note that $sFDR^\ell$ admits an equivalent recursive definition proceeding from level $\ell$ to level 0. Assign to each rejected hypothesis at level $\ell$ an error measure which is the type I error indicator for this hypothesis. Now, starting from level $\ell − 1$ and continuing to coarser
resolution levels, each selected parent hypothesis is assigned an error measure which is the average of the error measures of all the selected hypotheses in the family it indexes, until one reaches the root of the tree, $H_0$. Then $s\text{FDR}^\ell$ is the expectation of the error measure assigned to the root hypothesis. Formally,

$$s\text{FDR}^\ell = \mathbb{E} (\bar{C}_{i_0}).$$

This illustrates that for $\ell = 1$, $s\text{FDR}^\ell$ is the simple FDR for $F_1$, and for $\ell > 1$ the error measure can be viewed as a generalization of FDR for a tree of hypotheses, capitalizing on the selected families of hypotheses. Similar to the FDR, which is adaptive to the amount of signal within a family, becoming less stringent when there are many effects, the $s\text{FDR}^\ell$ is adaptive to the amount of signal in the selected families at level $\ell$ and their ancestor families.

Remark 3 We discussed in Section 3 the connection between $s\text{FDR}^1$, $s\text{FDR}^2$ and $s\text{FDR}^3_T$. We now generalize this claim for $s\text{FDR}^\ell$ for a tree with $L$ levels.

Proposition 4.1. Assume that for a given $\ell \in \{2, \ldots, L\}$, the hierarchical testing procedure guarantees that if a parent hypothesis at level $\ell - 1$ is rejected, at least one of the hypotheses within the family it indexes is rejected, i.e. for each selected $H_i$ at level $\ell - 1$, $S_i^{\ell-1}$ is not empty. Then $s\text{FDR}^{\ell-1} \leq s\text{FDR}^\ell$, i.e. control of $s\text{FDR}^\ell$ guarantees control of $s\text{FDR}^{\ell-1}$.

The proof of this proposition follows from the recursive definition of $s\text{FDR}^\ell$ above, and is deferred to the Appendix. This proposition implies that if the hierarchical testing procedure guarantees for all the levels $\ell = 1, \ldots, L - 1$, that each rejected parent hypothesis has at least one rejected child hypothesis, then control of $s\text{FDR}^\ell$ implies control of $s\text{FDR}^k$ for all $k = 1, \ldots, \ell - 1$. In particular, controlling $s\text{FDR}^L$ guarantees control of $s\text{FDR}^\ell$ for all the levels of the tree, so that one has confidence regarding the errors within the selected families in each level of the tree.

4.2 Testing strategy

We next present a general hierarchical testing strategy, which we refer to as TreeBH, targeting control of $s\text{FDR}^\ell$ at level $q_0^{(\ell)}$ for all $\ell \in \{1, \ldots, L\}$. We assume that one has the $p$-values for all the hypotheses in the tree. As discussed before, one may have only the $p$-values for the finest level hypotheses, and compute the $p$-values for hypotheses at
the coarser levels by combining the \( p \)-values within the families using Simes method, or using any other valid method for computing the global null \( p \)-value. Note that if one

**TreeBH: selection-adjusted hierarchical testing procedure**

**Input**: The target levels for error rates for sFDR, \( \ell = 1, \ldots, L \): \( q_0^{(1)}, \ldots, q_0^{(L)} \)

The \( p \)-values for all the hypotheses in the tree

```
begin
    S^0 ← i_0
    q_{i_0} = q_0^{(1)}
    for \( \ell = 1, \ldots, L \) do
        S^{\ell} ← ∅
    end
    for \( \ell = 1, \ldots, L \) do
        while \( S^{\ell-1} \neq ∅ \) do
            for \( i \in S^{\ell-1} \) do
                Apply the BH procedure at level \( q_i \) on the \( p \)-values of family \( F_i^{\ell} \)
                \( S^{\ell} ← S^{\ell} \cup S_i^{\ell} \)
                for \( j \in S_i^{\ell} \) do
                    \( q_j ← q_0^{(\ell+1)} \times q_i / q_0^{(\ell)} \times |S_i^{\ell}| / |F_i^{\ell}| \)
                end
            end
        end
end
```

**Output**: The set of all the rejected hypotheses, \( \bigcup_{\ell=1}^{L} S^{\ell} \).

is interested in discoveries only up to some specific level of resolution, say Level \( k < L \), then one should apply the TreeBH procedure to the subtree consisting of the first \( k \) levels of the original tree, i.e. one should replace \( L \) by \( k \) in the algorithm. Moreover, if one is interested only in discoveries within families at some specific levels, one may apply any selection rules for selecting hypotheses within the preceding levels (possibly based on the \( p \)-values in the families they index), leading to selection of families within the levels of interest. Only the families at those levels should be tested at the adjusted level corresponding to the number of selected hypotheses in the preceding levels. Finally, as we discussed above, the error rate sFDR can be modified so that different error measures are assigned to different families. If \( C_i^{(\ell)} \) is the error measure assigned to \( F_i^{\ell} \), then one should apply \( \mathbb{E}(C_i^{(\ell)}) \)-controlling procedure (instead of the BH procedure) at level \( q_i \) on the \( p \)-values of \( F_i^{\ell} \) in each step of the algorithm. Below we give the theoretical result
for the TreeBH procedure. The generalizations discussed above and the corresponding theoretical results are given in the Appendix.

4.3 Error rate control

To state the conditions under which we can prove that TreeBH controls the target error rate, it is useful to resort to the notion of ancestor hypotheses: the ancestor hypotheses of $F_{i}^{\ell}$ are $H_{i}$, the parent of $H_{i}$, and so on, until we reach the ancestor at Level 1. We can now introduce the following two assumptions:

A1 The BH procedure is valid for the dependence among the $p$-values within each family, e.g. the $p$-values are independent or satisfy the positive regression dependence on the subset of true null hypotheses property (PRDS, see [17]).

A2 For each level $\ell \in \{2, \ldots, L\}$ and each family $F_{i}^{\ell}$, the $p$-values of the hypotheses in $F_{i}^{\ell}$ are independent of the $p$-values of the hypotheses at levels $1, \ldots, \ell - 1$ which are not ancestors of the family $F_{i}^{\ell}$.

Note that the two assumptions above do not imply any dependency structure between the $p$-values of hypotheses in the same branch of the tree, i.e. ancestor hypotheses of each $H_{i}$ residing at level $L$. Finally, we would like to note that in the case we focus on, where each parent hypothesis is the intersection of all the hypotheses in the family it indexes (i.e. the global null), the $p$-value for the parent hypothesis will typically depend only on the $p$-values of the hypotheses in the family it indexes. Obviously this will happen when the $p$-value for the global null is the combination of $p$-values within its family (e.g. using Simes or Fisher’s method). In this case assumption A2 holds when the $p$-values in each family at level $L$ are independent of the $p$-values in any other family at level $L$, as we assumed in Theorem [1].

**Theorem 2.** If assumptions A1 and A2 hold, the TreeBH procedure with input parameters $(q(1), \ldots, q(L))$, guarantees for each $\ell \in \{1, \ldots, L\}$ that $sFDR_{\ell} \leq q(\ell)$.

This result follows from the theorem given in the Appendix, which addresses a more general procedure, addressing different error rates and selection rules for selecting the families.

5 Examples and simulations

In this section we consider examples and simulations with the intention of (1) illustrating the differences between the error rates and procedures we introduced with other
methods in the literature; (2) studying the effect of a type of dependence between the test statistics at Level 3 similar to that we expect in genetic studies.

In the interest of simplicity, we limit ourselves to \( L = 3 \) and \( m_i = m \) and \( k_{ij} = k_i \): we can describe the configurations of true and false nulls using a matrix containing all the Level 3 hypotheses, as in Table 1. Each row includes all the hypotheses \( H_{ijt} \) corresponding to one value of \( i \) (so that the presence of a non-null hypothesis in row \( i \) signifies that \( H_{i••} \) is false). Each column corresponds to one pair \((j, t)\), with all the columns with the same value \( j \) adjacent: so that, within a row, blocks of columns correspond to the families in Level 3, and the presence of a non-null hypothesis in row \( i \), block \( j \) signifies that \( H_{ij•} \) is false.

To compare the performance of different approaches we rely on level-specific error rates and power. Specifically, we calculate the FDR for discoveries at Levels 1, 2, and 3, which we denote by \( \text{FDR}^\ell \) for \( \ell = 1, 2, 3 \) respectively, as well as the selective sFDR for levels \( \ell = 2, 3 \). Note that sFDR\(^1\) is omitted since \( \text{FDR}^1 = \text{sFDR}^1 \). For the hypotheses at each level, we also calculate power. Across the different simulations, we compare our procedure with BH (as a standard of FDR control without reference to sub-groups of hypotheses) and p-filter which, while having a different target, shares many commonalities with our procedure (attempting to control the FDR for "group level discoveries" and relying on Simes’ combination \( p \)-value). Specifically, we use the following approaches and labels.

**BH** = Benjamini-Hochberg method \([1]\) applied across the pooled set of \( p \)-values for the entire matrix of hypotheses. This guarantees control of FDR\(^3\).

**BB** = Benjamini-Bogomolov method \([7]\) applied with hypotheses grouped into a two-level hierarchy with \( H_{i••} \) in Level 1, each indexing a family \( \mathcal{F}_i^2 = \{H_{ijt}, j = 1, \ldots m, t = 1, \ldots, k_j\} \). The selection in level 1 is done using BH on Simes’ \( p \)-values for \( H_{i••} \). This guarantees control of FDR\(^3\), as well as of a selective error rate on these second layer hypotheses (which we do not calculate here).

**p-filter – non-hier** = p-filter applied to the matrix of hypotheses in Table 1 and groups
defined by the pooled set of all hypotheses, rows, and column. This guarantees control of FDR$^1$ and FDR$^3$ and is considered for ease of comparison with [11].

\textbf{pfilter – hier} = p-filter applied with groups defined by the pooled set of all hypotheses, rows, and sets of columns (i.e. a nested setup mimicking our hierarchical procedures). This guarantees control of FDR$^\ell$ for $\ell = 1, 2, 3$.

\textbf{TreeBH} = the 3-level TreeBH, guaranteeing control of FDR$^1$, sFDR$^2$ and sFDR$^3$.

\textbf{R-BB} = the Repeated BB, guaranteeing control of FDR$^1$, sFDR$^2$ and sFDR$^3$.

5.1 Example: the differences across error rates and procedures

We start with a small example that allows us to explicitly work out the differences between the various error rates and procedures. We consider six hypotheses at Level 1, each indexing a family of 6 hypotheses, parents to families at Level 3 that contain 2,2,2,2,2, and 90 hypotheses respectively, with truth assignment as described in Table 2.

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

Table 2: The setting for Example 5.1. The non-null hypotheses are marked in red.

First we discuss the implications of the configuration in Table 2 on error rates. We note that 5 out of 6 Level 1 hypotheses are false, so we can expect FDR$^1$ will be contained for any method. Consider now the error control for Level 3 discoveries: methods such as BH and p-filter, which do not have consideration for families, will weight any false discovery against the many possible true discoveries in family $\mathcal{F}_{16}^3$ (the family which contains 90 non-null hypotheses). For the selective methods we propose, instead, any false discovery in families $\mathcal{F}_{\ell}^3$ for $\ell = 3, \ldots, 6$ would result in FDP$_{16}^3 = 1$ (and FDP$_{26}^3 \geq 0.5$) and this would contribute with a substantial weight to the average in sFDR$^3$.

We now consider how the power of the different procedures is influenced by the configuration in Table 2. A large number of the Level 3 families are homogeneous: this presents an advantage for testing procedures that recognize the families, thereby allowing the BH threshold for significance to adapt to the different proportions of non-null hypotheses.

Figure 2 reports the results of a simulation where all the error rates are targeted at level 0.1 and, for each realization, the p-values for each of the hypotheses are generated
independently as follows:

\[ X \sim \mu + \mathcal{N}(0, 1) \]

\[ \text{p-value} = 1 - \Phi(X), \]

where \( \Phi \) denotes the standard normal cdf, \( \mu = 0 \) for null hypotheses, and \( \mu > 0 \) for non-null hypotheses, where larger values of \( \mu \) correspond to greater signal strength. Figure

Figure 2: Results for Example 5.1. Each point corresponds to the average of 1000 realizations. Dashed horizontal lines indicate the target values for the error rates.

underscores how each of the methods controls the designed target error rates, but not others. BH does not control any Level 1 or Level 2 error, nor sFDR\(^3\), and the p-filter methods do not control all of the sFDR\(\ell\). In this set-up it appears that BB controls the sFDR\(\ell\), but we will see in other examples that this is not always the case. Figure 2 also shows how the TreeBH and R-BB procedures, despite showing the most stringent error control in this example, have the highest power across levels (beaten only by BH in Level 1, where this procedure has no FDR control). Interestingly, their power in Level 2 and 3 is higher than that of BB, which shares some of the 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, as anticipated.
Higher power at Level 1 is due to the fact that in calculating the Simes’ $p$-values for $H_i$, one
uses 6 $p$-values (rather than 100), and, at least for $i = 3, 4, 5$, five of those are going
to be very small, due to the non null status of the hypotheses they represent: the effect
of these $p$-values would be more washed out in the entire pool of 100 hypotheses.

5.2 Example: the p-filter set-up

For ease of comparison with the literature, we consider here the same example described
in Section 6.2 (“Multilayer setting”) of [11]. There are a total of $100 \times 100$ hypotheses, and,
as illustrated in the upper left panel of Figure 3, the non-null hypotheses (true signals)
are arranged into 2 blocks of size $15 \times 15$ with 15 additional non-null hypotheses along
the diagonal. In the simulations, p-values are generated using the same rule described
in section 5.1.

Figure 3: True signals (marked in blue in panel labelled “truth”) alongside results for an example
run for all methods compared. The methods in the top row either do not group the hypotheses at
all (BH), use a hierarchical grouping by row (BB), or group both by row and column (p-filter -
row/col). The methods in the bottom row use the column groupings demarcated with dashed lines
in the lower left panel to group the hypotheses within each row.

A single illustrative result using $\mu = 3$ is given in Figure 3 and a full performance
comparison across a range of \( \mu \) values i.e. for varying signal strength, averaged over 100 iterations, is given in Figure 4. All methods are applied with target error rate 0.2. These results demonstrate that standard BH drastically fails to control the Level 1 and Level 2 FDR and the selective error rates, while BB and p-filter - row/col, although they control the Level 1 FDR, do not control the Level 2 FDR or the Level 2 and Level 3 selective error rates. P-filter-hier is able to control all error rates considered, but achieves lower Level 3 power vs. the TreeBH and R-BB methods. The TreeBH and R-BB, on the other hand, offer the best overall performance in terms of error control and power, even if many of the differences in this example are small.

![Figure 4: Results for Example 5.2, the “p-filter set-up”. Each point corresponds to the average of 100 realizations. Dashed horizontal lines indicate the target values for the error rates.](image)

5.3 Multi-trait GWAS

Having clarified the conceptual differences between our proposal and existing methods, we now explore the properties of TreeBH and R-BB in a set-up that resembles multi-trait genome-wide association studies (GWAS): our goal is to investigate the effects of
signal sparsity and of the dependence that exists among different tests. We adapt the simulation study proposed in Lewin et al. (2015) [18] for eQTL studies. Specifically, we assume the same underlying model, with \( X_{n \times p} \) the matrix of additively coded values for each of \( p \) genetic variants in \( n \) subjects. The \( n \times q \) matrix \( Y_l \) represents the simulated gene expression values for tissue \( l \), \( 1 \leq l \leq L \), and is generated from the \( X \) matrix via the linear model

\[
Y_l = XB_l + E_l + E_{\text{shared}},
\]

where \( B \) is a matrix of regression coefficients which is shared across tissues, \( E_l \) is a matrix of tissue-specific residuals with entries \( e_{ikl} \sim \mathcal{N}(0, \sigma^2_l) \), and \( E_{\text{shared}} \) is a matrix of residuals shared across tissues with entries \( e_{ik\text{shared}} \sim \mathcal{N}(0, \sigma^2_{\text{shared}}) \). Following [18], we assume the same pattern of association within each tissue (i.e. the coefficient matrix \( B_l = B \) is common across tissues). Note that this implies that the Level 3 families are homogenous. The total variance of the residuals for tissue \( l \) is then \( \sigma^2_{\text{total}} = \sigma^2_{\text{shared}} + \sigma^2_l \). The entries in \( B \) are generated following the equation \( \beta_{kj} = \lambda_{kj} \times \gamma_{kj} \), where the magnitude of the coefficients \( \lambda_{kj} \sim \mathcal{N}(\mu, 0.001^2) \) is controlled by the signal strength parameter \( \mu \) and \( \gamma_{kj} \) is a binary indicator (i.e. \( \gamma_{kj} \) determines the pattern of association, while \( \lambda_{kj} \) controls the signal strength).

To better mimic a real multi-tissue eQTL study, we increase the dimensionality vs. that of [18] (who consider a rat dataset with \( n = 29 \) and \( p = 1304 \) as their predictors and simulate the expression of \( q = 150 \) genes in \( L = 3 \) tissues), and take \( X_{n \times p} \) to be the chromosome 17 genotype data from the NFBC data set (\( n = 5402 \) and \( p = 8713 \)), where missing SNP values are replaced by the column mean and the data have been standardized. Given this genotype matrix, we then simulate the expression of \( q = 250 \) genes in each of \( L = 5 \) tissues following equation (5).

Following the scenario in [18] with correlated noise, balanced across tissues, we set \( \sigma^2_l = 0.0084 \) and \( \sigma^2_{\text{shared}} = 0.0016 \), resulting in total noise standard deviation \( \sigma_{\text{total}} = 0.1 \). Varying signal-to-noise ratios are achieved by varying the signal strength parameter \( \mu \) (i.e. the mean of the normal distribution from which the non-zero entries in \( B \) are drawn). We assume that there are 50 causal SNPs which are each associated to the expression of 5, 10, or 20 genes. The “true” SNPs are selected at random in each iteration, but the configuration of which phenotype the “true” SNPs affect are held constant. This configuration is illustrated in Figure S1.

All \( p = 8713 \) SNPs are tested for association to the simulated gene expression traits. Association \( p \)-values are obtained using Matrix eQTL [15], including the first 5 principal components of the genotypes as covariates to account for population structure. Discoveries are considered to be true positives as long as the discovered SNP is within 1Mb and has correlation of magnitude at least 0.2 with a SNP that is truly relevant to the trait. Given this definition, we compare the performance of BH, BB, R-BB, and TreeBH, where
BB is applied with SNPs in Level 1 and all traits (i.e. expression for all genes across all tissues) in Level 2, and R-BB and TreeBH are applied with SNPs in Level 1, genes in Level 2, and tissues in Level 3. The p-filter is not included here as it was not able to run in a timely manner (i.e. < 24 hrs) given similar groupings to those in the simulation with independent hypotheses. All methods are applied with target level 0.05. A performance comparison for the 4 methods compared across 25 simulated data sets is given in Figure 5: the TreeBH and R-BB procedures appear to at least approximately control the target error rates, unlike BH, and have a similar FDR, sFDR, and power to BB. In particular, we see that BB is close to controlling the target error rates, unlike in the simulations from Section 5.2. This is likely due to the fact that the adjustment for step 1 selection is substantially more stringent here because of the sparsity of signal across SNPs (specifically, the fact that only 50 of 8713 Level 1 hypotheses are non-null).

Figure 5: Results for Example 5.3: eQTL simulation. Each point corresponds to the average of 25 simulated data sets using the full set of $p = 8713$ SNPs on chromosome 17.
6 Case studies

6.1 Genetic regulation of gene expression across multiple tissues

The goal of expression quantitative trait loci (eQTL) analysis is to identify DNA variants (typically SNPs) that influence the expression of genes. Since gene expression levels differ across tissues, eQTL analysis may reveal both shared and tissue-specific patterns of regulation. In addition to providing insight into the architecture of genetic regulation across tissues, the findings of multi-tissue eQTL analysis can help pinpoint disease mechanisms by linking risk variants from genome-wide association studies (GWAS) to the genes they regulate in specific tissues. The problem of identifying relevant associations is challenging, however, given the large number of hypotheses under consideration and the degree of noise in the data. In this context, organizing family of hypotheses in a hierarchical structure is fairly natural.

We now provide some additional background on the motivating problem of eQTL analysis. In the eQTL setting, there are typically assumed to be two classes of regulation: local (in which variants affect the expression of nearby genes, typically assumed to be within 1Mb of the transcription start site of the gene), and distal (in which variants affect the expression of genes located far away on the genome, perhaps even on different chromosomes). Local association, which may reflect mechanisms such as the disruption of a transcription factor binding site, is much more common than distal regulation, and is believed to often be shared across multiple tissues. Given the very large number of hypotheses under consideration and the possibility of confounding factors, distal associations have proven to be very difficult to detect. It is therefore common to focus on the more highly powered setting of local regulation, and we follow here this approach.

Typically, the $H_{ijt}$ hypotheses are tested separately for each tissue using a simple linear model relating gene expression to genetic variation. The collection of hypotheses $\{H_{ijt}\}$ can be organized in different ways. The most common approach in the eQTL literature has been to perform error control in each tissue separately [19, 20]. Results across tissues are then compared and conclusions are drawn on the tissue-specific nature of the detected gene-SNP associations. It has been noted that this approach is error prone and that joint analysis of multiple tissues is likely to result in lower false positives and false negatives. Methodology based on meta-analysis [21] and on Bayesian model selection [22, 23] has been proposed 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.

The genotype-tissue expression (GTEx) project is an NIH-funded project that aims to characterize variation in gene expression and genetic regulation across tissues. The Phase 1 data release includes 450 subjects and 44 tissues with at least 60 samples. In the
Phase 1 data release, gene expression was measured for around 21,000–34,000 genes per tissue, and genotypes were estimated for around 11 million SNPs. As in the simulation using real genotypes, we focus on the set of reasonably independent SNPs filtered to have local $R^2 < 0.5$. After tissue-specific QC, this set includes between 250,000 and 300,000 SNPs per tissue for each of the 44 tissues, with a total of 305,820 SNPs passing QC in at least one tissue.

For each tissue, the $H_{ijt}$ hypotheses are tested using a simple linear model with normalized expression for gene $j$ as the response and the estimated number of copies of the minor allele for SNP $i$ as the predictor. Gender, array platform, PEER factors (to adjust for confounding factors affecting global levels of gene expression), and genotype principal components (to adjust for population structure) are included as covariates. The corresponding $p$-values $p_{ijt}$ are generated using the software FastQTL [16].

One possible finding of interest in multi-tissue eQTL analysis is a set of eSNPs i.e. SNPs which we believe to play a functional regulatory role in at least one tissue. This type of discovery is of particular interest when eQTL results are used to infer the effect of the studied genetic variants on other phenotypes: eSNPs are known to have some functional effect and are therefore more likely to have an impact on other phenotypes as well. Given this goal, we could naturally group the hypotheses into a hierarchical structure with SNPs in Level 1, genes in Level 2, and tissues in Level 3. The Level 1 hypothesis $H_{i\bullet \bullet}$ addresses the question: does SNP $i$ have effects on expression in any tissue? The corresponding null hypothesis $H_{i\bullet \bullet}$ corresponds to SNP $i$ having no gene associations in any tissue. We consider SNP $i$ to be an eSNP if we reject $H_{i\bullet \bullet}$, and a SNP-gene pair to be discovered if we reject $H_{ij\bullet}$. $P$-values are defined starting from the leaf hypotheses, which receive the $p$-value calculated by FastQTL. $P$-values for the global nulls corresponding to the sets of SNP×gene hypotheses across tissues are defined using Simes, and $p$-values for the global nulls corresponding to the SNPs are defined with Simes again.

In the multi-tissue eQTL context, the TreeBH procedure controls then the following error rates: the FDR for eSNPs; the expected average proportion of false SNP-gene associations across the selected SNPs; and a weighted average of proportion of false tissue discoveries for the selected SNP-gene pairs. We note that other ways of organizing hypotheses in a tree are also meaningful in eQTL studies, for example, in our participation in the GTEx consortium we have used other hierarchical structures, when the main discovery of interest are eGenes. We focus here on the tree described above, as it resembles more generally the type of hierarchical structure that is meaningful in multi-trait association studies, and it allows us to annotate SNPs meaningfully.

To provide a benchmark for our results, in addition to TreeBH we analyzed the data with two other approaches: the BH procedure applied to the SNP-gene association $p$-
values separately by tissue ("BH sep"), and the BH procedure applied to the pooled set of p-values from all tissues ("BH pooled"). The total number of discoveries for each of the adopted procedures is reported in Table 3. The eSNP percentages show that a considerable proportion of the SNPs tested for local association are selected across all methods. The number selected for BH pooled is only slightly less than the number selected for BH sep. This is due to the fact that the BH threshold is adaptive, and in this setting, the proportion of small p-values is quite substantial i.e. there is a large proportion of SNP-gene-tissue association hypotheses that are non-null. The TreeBH procedure is much more conservative at the SNP level (because it provides control of the eSNP FDR), but relatively less stringent in selecting the genes and tissues for these eSNPs, resulting in a similar number of genes associated to each eSNP as for the BH methods, 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 1 tissue. Given the hypothesis that local regulatory relationships are likely to be shared across tissues, the TreeBH results seem more biologically plausible vs. the results from the BH methods.

In Figure 6, we provide a more detailed look at the sharing of SNP-gene pairs across all 44 tissues, the 10 brain tissues, and the remaining 34 non-brain tissues. We see that the proportion of SNP-gene pairs found in 1 tissue only had a larger decline when comparing the results for BH sep vs. TreeBH for the set of 10 brain tissues rather than all tissues. Specifically, the proportion of SNP-gene pairs discovered in 1 tissue only in all tissues was 0.61 for BH sep vs. 0.48 for TreeBH (21% fewer), while in brain tissue the

| Level  | # eSNPs | BH sep | BH 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%    |

Table 3: Numerical comparison of selection results for the BH procedure applied separately by tissue ("BH sep"), BH procedure applied to the pooled set of p-values from all tissues ("BH pooled"), and from the 3-level TreeBH procedure ("TreeBH"). "%eSNPs" is the percentage of SNPs selected out of the 305,820 SNPs tested for association. "# genes per eSNP" and "# tissues per SNP-gene pair" are averages across all discovered eSNPs and all discovered SNP-gene pairs, respectively. "% SNP-gene pairs 1 tissue only" is the percentage of discovered SNP-gene pairs that were associated in only one tissue.
proportion was 0.58 for BH sep vs. 0.41 for TreeBH (29% fewer). This suggests that the hierarchical method is in fact capitalizing on shared signal across closely related tissues.

The converse of this result is illustrated in Figure 7, which reports the number of SNP-gene pairs discovered in one tissue only (i.e. tissue-specific discoveries). While all tissues have fewer tissue-specific discoveries under the TreeBH procedure vs. separate BH, some tissues (in particular, Testis) retain a large number of tissue-specific associations.

![Figure 6: Comparison of selection results for the BH procedure applied separately by tissue ("BH sep") and the 3-level TreeBH procedure ("TreeBH") in terms of sharing of SNP-gene pair discoveries across tissues for all 44 tissues, the 34 non-brain tissues, and the 10 brain tissues.]

### 6.2 Association of the gut microbiome to colorectal cancer

The data considered here were originally collected as part of study examining the association of gut microorganisms with colorectal cancer [24], and are provided in the R package phyloseq [25]. For full details on the sample collection and processing, see [24]. Briefly, to quantify the relative abundance of microbial species, 16S rDNA (a variable region of the bacterial and archaeal genome which can be used for taxonomic classification) was amplified using PCR and then read via pyrosequencing. The resulting data are summarized as a table of counts for each operational taxonomic unit (OTU, a grouping of microorganisms similar to species but based on DNA sequence similarity) in each sample.

Here we consider the $n = 177$ samples with sufficient sample quantity (at least 500 total reads) and available diagnosis information (86 tumor, 91 normal colon). We focus on the $p = 496$ bacterial OTUs that are present in at least 10% of samples. We compute $p$-values of difference between tumor and normal samples for each OTU using the negative
Figure 7: Comparison of selection results for the BH procedure applied separately by tissue (transparent) and the 3-level TreeBH procedure (solid) in terms of the number of tissue-specific SNP-gene pairs per tissue i.e. the number of SNP-gene pairs that were discovered only in the given tissue.

The binomial model for differential expression implemented in DESeq2 [26]. These OTUs are related within a taxonomic tree with the following levels: Kingdom - Phylum - Class - Order - Family - Genus - Species. All OTUs considered here are part of the kingdom bacteria.

We are interested in controlling error rates at multiple levels in the taxonomic tree since the results of microbiome analysis are often discussed in terms of the taxonomic groupings discovered (e.g. phyla or genera) after controlling the FDR across OTUs. There is also an understanding that microbes which are closely related within a taxonomic tree are likely to have similar functions and roles in human disease.

For the purpose of applying the TreeBH procedure, OTUs with missing taxonomic information (specifically, a missing value for order, family, genus, or species) are grouped together as the category "Unknown" (nested within the appropriate hierarchy). The selections obtained from applying the TreeBH procedure with 8 levels (Kingdom - Phylum - Class - Order - Family - Genus - Species - OTU) using $q = 0.05$ in each level of the tree are shown in Figure 8. Specifically, the TreeBH procedure results in the selection of 6 phyla, 8 classes, 8 orders, 13 families, 15 genera, 17 species, and 19 OTUs. The number of selected OTUs is larger than the number of selected species as 2 "Unknown" species are each represented by 2 OTUs. Applying the Benjamini-Hochberg procedure across the $p$-values for all 496 OTUs at level 0.05 results in the selection of 33 OTUs, so
the hierarchical procedure is more conservative in terms of the raw number of OTUs discovered. The hierarchical procedure does, however, include 4 additional species not discovered using BH: in particular, 2 additional species of Fusobacteria are discovered, which are key species of interest given the focus on the genus *Fusobacterium* in [24], which was titled "Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma". Another of the additional species discovered (*Akkermansia muciniphila*) has been previously associated with colorectal cancer specifically, and was reported as one of the primary findings of [27]. The final additional species discovery under the hierarchical method (*Gemella haemolysans*) has been previously associated to oral cancer in a study of the oral microbiome [28]. 14 of the 15 species discovered under BH which were not discovered under TreeBH were categorized as "Unknown". Thus, at a high level, the results of the TreeBH method, while more conservative overall, were enhanced in terms of biologically interpretable findings.

**Conclusions**

We have introduced a novel hierarchical procedure that allows the selective testing of families of hypotheses that appear to hold more promise for discoveries, potentially increasing power without inflating error rates.

The TreeBH and Repeated BB procedures guarantee control of selective level-specific error rates under certain assumptions on the dependency structure of p-values within the tree. Our simulations show how controlling these newly defined selective error rates often implies control of other level-specific error measures. Furthermore, we have shown that in scenarios designed to mimic the dependence and sparsity structure of multi-trait GWAS, the procedures continue to control (at least approximately) the target error rates.

We also showed that our procedure can be modified to control other selective error rates that are not based on FDR, just as the results in [7] hold for a variety of error rates that can be expressed as averages. This makes the hierarchical testing approach we present quite flexible, and adaptable to multiple settings, where the number of hypotheses in families at some levels, for example, is small enough to make controlling the FDR an unsatisfactory goal.

The TreeBH procedure we describe has many points of contact with the p-filter approach put forward in [11]: in both cases one assumes that p-values for the finer scale hypotheses are available, and these are summarized with Simes method to obtain an evaluation of the strength of group-level hypotheses. When adapted to our hierarchical organization of hypotheses, p-filter controls the level-restricted FDR, but has no consideration for the distinct families that make up the entire collection of hypotheses at a given level. Therefore, it does not control the selective error rates that we define. Fur-
Figure 8: Taxonomic tree of selections using TreeBH procedure. Additional discoveries from TreeBH not found using BH are bolded and marked with a dashed line.
thermore, it is impossible for p-filter to gain power by adapting to the different sparsity levels across families. Finally, it is worth noting how the computational time required by p-filter is substantially higher than that of TreeBH, making its application in genomic problems more difficult. We note that in a recent preprint [12], the p-filter framework has been extended to control FDR for any identified family of hypotheses: the set of families for which one desires this control has, however, to be specified in advance, rather than selected on the basis of the data. Technically it is possible within this framework to control FDR for each family, however this strategy does not guarantee control of our selective error rates, because it does not adjust for the data-based selection of families. This is demonstrated in [7] for a tree with two levels.

Our procedures are extensions of the proposals in [7] and therefore have some of the same merits and limitations. In particular, we want to point out that a recent work [10] underscores how selective error rates might be controlled with higher power when a conditional approach to testing is possible. They demonstrate the feasibility for two layers structure and under independence at the second level. In principle, as long as exact conditional distributions can be evaluated, it might be possible to adopt the conditional testing approach in [10] to also control the selective error rates that we introduced here. Currently it is not obvious to us how it can be done in a general hierarchical structure.

The study of eQTLs across multiple tissues presents such a formidable multiplicity challenge that it has motivated not only us, but also others to investigate novel procedures. Among the most recent developments we would like to single out [29], which takes an empirical Bayes approach with substantial possible power gains, and whose analysis of GTEx data has informed some of the displays we presented. Differently from the view-point we adopt, however, [29] does not offer control of the FDR for global discoveries.

**Acknowledgements**

The authors would like to thank the donors that made possible the GTEx data collection and all the scientists that participated in the consortia for sharing data and insights. Y.B. and C.S. were partially supported by NIH grant HG006695; C.P. and C.S. were partially supported by NIH grant MH101782; C.P. was partially supported by NIH/NCI grant P30CA016672; M.B. was partially supported by the Israel Science Foundation grant no. 1112/14. We acknowledge use of the dataset “STAMPEED: Northern Finland Birth Cohort 1966” (phs000276.v2.p1) available via dbGap.
References

[1] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society, Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[2] M. Perone Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman, “False discovery control for random fields,” *J. Amer. Statist. Assoc.*, vol. 99, no. 468, pp. 1002–1014, 2004.

[3] Y. Benjamini and R. Heller, “False discovery rates for spatial signals,” *J. Amer. Statist. Assoc.*, vol. 102, no. 480, pp. 1272–1281, 2007.

[4] W. Sun, B. J. Reich, T. T. Cai, M. Guindani, and A. Schwartzman, “False Discovery Control in Large-Scale Spatial Multiple Testing,” *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 77, pp. 59–83, Jan 2015. PMID: 25642138

[5] D. O. Siegmund, B. Yakir, and N. Zhang, “The false discovery rate for scan statistics,” *Biometrika*, vol. 98, pp. 979–985, 2011.

[6] D. Brzyski, C. B. Peterson, P. Sobczyk, E. J. Candes, M. Bogdan, and C. Sabatti, “Controlling the rate of GWAS false discoveries,” *Genetics*, vol. 205, no. 1, pp. 61–75, 2017.

[7] Y. Benjamini and M. Bogomolov, “Selective inference on multiple families of hypotheses,” *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 76, no. 1, pp. 297–318, 2014.

[8] C. B. Peterson, M. Bogomlov, Y. Benjamini, and C. Sabatti, “Many Phenotypes Without Many False Discoveries: Error Controlling Strategies for Multitrait Association Studies,” *Genet. Epidemiol.*, vol. 40, pp. 45–56, Jan 2016. PMID: 26626037

[9] C. B. Peterson, M. Bogomlov, Y. Benjamini, and C. Sabatti, “TreeQTL: hierarchical error control for eQTL findings,” *Bioinformatics*, vol. 32, no. 16, pp. 2556–2558, 2016. PMID: 27153635

[10] R. Heller, N. Chatterjee, A. Krieger, and J. Shi, “Post-selection inference following aggregate level hypothesis testing in large scale genomic data,” *bioRxiv*, 2016.

[11] R. Foygel Barber and A. Ramdas, “The p-filter: multi-layer false discovery rate control for grouped hypotheses,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Nov. 2016.
[12] A. Ramdas, R. Foygel Barber, M. J. Wainwright, and M. I. Jordan, “A Unified Treatment of Multiple Testing with Prior Knowledge,” ArXiv e-prints, Mar. 2017.

[13] D. Yekutieli, “Hierarchical false discovery rate-controlling methodology,” J. Amer. Statist. Assoc., vol. 103, no. 481, pp. 309–316, 2008.

[14] R. J. Simes, “An improved Bonferroni procedure for multiple tests of significance,” Biometrika, vol. 73, no. 3, pp. 751–754, 1986.

[15] A. A. Shabalin, “Matrix eQTL: ultra fast eQTL analysis via large matrix operations,” Bioinformatics, vol. 28, no. 10, pp. 1353–1358, 2012.

[16] H. Ongen, A. Buil, A. B. Brown, E. T. Dermitzakis, and O. Delaneau, “Fast and efficient QTL mapper for thousands of molecular phenotypes,” Bioinformatics, vol. 32, no. 10, pp. 1479–1485, 2015. PMID: doi:10.1093/bioinformatics/btv722.

[17] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency,” Annals of Statistics, vol. 29, no. 4, pp. 1165–1188, 2001.

[18] A. Lewin, H. Saadi, J. E. Peters, A. Moreno-Moral, J. C. Lee, K. G. Smith, E. Petretto, L. Bottolo, and S. Richardson, “MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues,” Bioinformatics, vol. 32, no. 4, pp. 523–532, 2016.

[19] A. C. Nica, L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, et al., “The architecture of gene regulatory variation across multiple human tissues: the MuTHER study,” PLoS Genetics, vol. 7, no. 2, p. e1002003, 2011.

[20] E. Grundberg, K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, et al., “Mapping cis- and trans- regulatory effects across multiple tissues in twins,” Nature Genetics, vol. 44, no. 10, pp. 1084–1089, 2012.

[21] J. H. Sul, B. Han, C. Ye, T. Choi, and E. Eskin, “Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches,” PLoS Genet, vol. 9, no. 6, p. e1003491, 2013.

[22] G. Li, A. A. Shabalin, I. Rusyn, F. A. Wright, and A. B. Nobel, “An empirical Bayes approach for multiple tissue eQTL analysis,” ArXiv e-prints, Nov. 2013.

[23] T. Flutre, X. Wen, J. Pritchard, and M. Stephens, “A statistical framework for joint eQTL analysis in multiple tissues,” PLoS Genet., vol. 9, p. e1003486, May 2013.
Appendix

Proof of Proposition 2.1

In order to prove the proposition, we show that each of the selective error rates above can be decomposed into two or more terms, where each term addresses the errors within the selected families at a specific level. Let us denote by $I_0^1$ and $I_1^1$ the indices of true null hypotheses at Level 1 and the indices of false null hypotheses at Level 1, respectively, so that $I_0^1 \cup I_1^1 = \{1, \ldots, n\}$. Similarly, let us denote by $I_{0,i}^2$ and $I_{1,i}^2$ the indices of true null hypotheses and the indices of false null hypotheses within $\mathcal{F}_i^2$ respectively, for
Let us consider term (6). Note that if a family consists only of true null hypotheses, one or more rejections result in $FDP = 1$ within this family. Under condition 1, for each $i \in S^1$, at least one hypothesis in $F^2_i$ is rejected. Therefore $FDP_i = 1$ for $i \in S^1 \cap I^0_i$, so

$$
\mathbb{E} \left( \frac{1}{\max\{|S^1|, 1\}} \sum_{i \in S^1} FDP_i \right) = \mathbb{E} \left( \frac{1}{\max\{|S^1|, 1\}} \sum_{i \in S^1 \cap I^0_i} FDP_i \right) + \mathbb{E} \left( \frac{1}{\max\{|S^1|, 1\}} \sum_{i \in S^1 \cap I^1_i} FDP_i \right).
$$

Let us now consider $sFDR^3_T$ under conditions 1 and 2, which is given by

$$
\mathbb{E} \left( \frac{1}{\max\{|S^1|, 1\}} \sum_{i \in S^1} \sum_{j \in S^2_i} \frac{1}{\max\{|S^2_i|, 1\}} FDP_{ij} \right) = \mathbb{E} \left( \frac{|S^1 \cap I^0_i|}{\max\{|S^1|, 1\}} \right),
$$

which is the Level 1 FDR. Thus term (6) is the Level 1 FDR addressing the errors at Level 1, while term (7) addresses the errors at Level 2 within the selected families which do not contain only true null hypotheses. The decomposition above shows that under condition 1, control of $sFDR^2_T$ implies control of $FDR^1$.
Thus we obtain

$$
\mathbb{E}\left(\frac{1}{\max\{|S^1|, 1\}} \sum_{i \in S^1} \sum_{j \in S^2 \cap I^2_{i}} \frac{1}{\max\{|S^2_i|, 1\}} FDP_{ij}\right) = \mathbb{E}\left(\frac{1}{\max\{|S^1|, 1\}} \sum_{i \in S^1} \frac{|S^2_i \cap I^2_{i,j}|}{\max\{|S^2_i|, 1\}} \right)
= \mathbb{E}\left(\frac{1}{\max\{|S^1|, 1\}} \sum_{i \in S^1} FDP_i \right).
$$

(10)

The expression in (10) is $sFDR^2$, therefore we obtain that term (8) is $sFDR^2$. It follows from the decomposition of $sFDR^3_T$ into the sum of (8) and (9), that under conditions 1 and 2 control of $sFDR^3_T$ implies control of $sFDR^2$. We have shown that under condition 1 $sFDR^2$ implies control of $FDR^1$. Therefore under conditions 1 and 2 control of $sFDR^3$ implies both control of $sFDR^2$ and control of $FDR^1$. Using the decomposition of $sFDR^3$ into the sum of $sFDR^2$ and (9), and the decomposition of $sFDR^2$ into the sum of (6) and (7), we obtain the following decomposition of $sFDR^3_T$:

$$
\mathbb{E}\left(\frac{1}{|S^1|} \sum_{i \in S^1} \sum_{j \in S^2_i} \frac{1}{\max\{|S^2_i|, 1\}} FDP_{ij}\right) = \mathbb{E}\left(\frac{|S^1 \cap I^1_0|}{\max\{|S^1|, 1\}} \right) + \mathbb{E}\left(\frac{1}{\max\{|S^1|, 1\}} \sum_{i \in S^1 \cap I^1_0} FDP_i \right) + \mathbb{E}\left(\frac{1}{\max\{|S^1|, 1\}} \sum_{i \in S^1} \sum_{j \in S^2_i \cap I^2_{i,j}} \frac{1}{\max\{|S^2_i|, 1\}} FDP_{ij}\right)
$$

(11)

(12)

(13)

This shows that under conditions 1 and 2, $sFDR^3_T$ consists of three terms, where (11) is the $FDR^1$ addressing the errors at Level 1, (12) addresses the errors at Level 2 within the selected families which do not consist only of true null hypotheses, and (13) addresses the errors at Level 3 within the selected families which do not consist only of true null hypotheses. Under condition 1, the sum of (11) and (12) is $sFDR^2$.

**Proof of Theorem 1**

Note that since the $p$-value for each parent hypothesis is a combination of the $p$-values for the hypotheses in the family it indexes, the assumptions of Theorem 1 on the dependence of $p$-values of Level 3 hypotheses implies that all the $p$-values for Level 2 hypotheses are jointly independent, and all the $p$-values for Level 1 hypotheses are jointly
independent. Therefore, the result in item 1 follows from the fact that the BH procedure is applied on the Level 1 $p$-values. The result in item 2 follows immediately from Theorem 1 in [7], since Step 2 makes the adjustment for selection at Step 1 according to the BB procedure. The result in item 3 follows from Theorem 2 for the general procedure. Finally, the result in item 4 follows from Theorem 1 in [7], when we consider steps 1 and 2 in the 3-level TreeBH procedure as a single selection rule of families in Level 3. It is easy to see that the assumptions of Theorem 1 in [7] hold: the assumptions on the dependency structure hold since the $p$-values in each Level 3 family are independent of $p$-values in any other Level 3 family, and the selection rule consisting of steps 1 and 2 is indeed simple, because BH procedure defines a simple selection rule (see [7] for a definition of a simple selection rule and examples).

Proof of Proposition 4.1.

Assume that for a given Level $\ell$, the procedure guarantees that for each rejected hypothesis at Level $\ell - 1$, at least one hypothesis within the family it indexes is rejected. We will show that the error measure assigned to each hypothesis at Level $\ell - 1$ by sFDR$^{\ell - 1}$ is smaller or equal than the error measure assigned to the same hypothesis by sFDR$^\ell$, when we use the recursive definition of sFDR$^\ell$'s. Let $H_i$ be a rejected hypothesis at Level $\ell - 1$, and let $\overline{C}_i^{\ell - 1}$ and $\overline{C}_i^\ell$ be its error measure according to sFDR$^{\ell - 1}$ and sFDR$^\ell$ respectively. For sFDR$^{\ell - 1}$, $\overline{C}_i = I_{[H_i \text{ is null}]}$, while for sFDR$^\ell$, $\overline{C}_i = \sum_{j \in S_i} I_{[H_j \text{ is null}]} / |S_i^\ell|$. If $H_i$ is a true null hypothesis, than its error measure according to sFDR$^{\ell - 1}$ is 1. On the other hand, in this case all the hypotheses in the family indexed by $H_i$ are also true, and at least one hypothesis in this family is rejected. This yields that the FDP in $F_i^\ell$ is also 1, i.e. $\sum_{j \in S_i} I_{[H_j \text{ is null}]} / |S_i^\ell| = 1$. Thus in the case where $H_i$ is a true null hypothesis, the error measure assigned to this hypothesis is 1 both for sFDR$^\ell$ and sFDR$^{\ell - 1}$. If $H_i$ is a false null hypothesis, the error measure assigned to it by sFDR$^{\ell - 1}$ is 0, while the error measure assigned to it by sFDR$^\ell$ is the FDP in $F_i^\ell$, which is 0 if no false null hypothesis within this family is rejected, and is positive otherwise. Thus we have proved that for each rejected hypothesis at Level $\ell - 1$, its error measure corresponding to sFDR$^{\ell - 1}$ is smaller or equal to that corresponding to sFDR$^\ell$. Since the error measure for each parent hypothesis is the average across the error measures of all its child hypotheses, and the final error measure corresponds to the error measure assigned to the root of the tree, we obtain immediately that sFDR$^{\ell - 1} \leq$ sFDR$^\ell$. 
Methodology and theoretical results for a general class of error rates

Similarly to [7], we consider a general class of error rates that can be written in the form $E(C)$ for some measure of errors $C$. These error rates include the family-wise error rate (FWER), $P(V \geq 1) = E(I_{V \geq 1})$, where $V$ is the number of type I errors, the FDR, and others (see [7] for additional examples). Assume that each family $\mathcal{F}_i^\ell$ is assigned with an error measure $C_i^\ell$. The general error rate we address is defined as follows.

$$E(sC) = \mathbb{E}\left(\sum_{i_0 \in S_0^0} 1 \sum_{i_1 \in S_1^0} \frac{1}{\max\{|S_1^0|, 1\}} \sum_{i_2 \in S_2^1} \frac{1}{\max\{|S_2^1|, 1\}} \cdots \sum_{i_{\ell-1} \in S_{\ell-2}^{\ell-1}} \frac{1}{\max\{|S_{\ell-1}^{\ell-1}|, 1\}} C_i^{\ell-1}\right),$$

or equivalently

For each selected hypothesis $H_i$ at level $\ell - 1$, $\overline{C}_i = C_i^{\ell}$.

For each selected hypothesis $H_j$ at level $k \in \{0, \ldots, \ell - 2\}$, $\overline{C}_j = \frac{\sum_{r \in S_{j+1}^k} \overline{C}_r}{|S_{j+1}^k|}$.

Then

$$E(sC) = E(\overline{C}_{i_0}),$$

where $H_{i_0}$ is the root node, i.e. the parent of $\mathcal{F}_1$ that is always selected. We next present a general hierarchical testing strategy for the case where the investigator is interested only in the discoveries in specific levels $\ell \in \mathcal{L}$, where $\mathcal{L} \subseteq \{1, \ldots, L\}$, targeting control of $E(sC)$ for all $\ell \in \mathcal{L}$. In this case one may use any selection rules for selecting the hypotheses within the families belonging to levels $\ell \notin \mathcal{L}$, since they only serve for selection of families residing at the levels of interest and are not interesting discoveries by themselves. Note that if the selection rules are such that if a parent hypothesis is selected, at least one of its child hypotheses is selected as well, Proposition 4.1 implies that one will obtain control of $E(sC)$ for all $k \leq \max\{\ell : \ell \in \mathcal{L}\}$. In order to prove that the general selection-adjusted hierarchical procedure (TreeC) controls $E(sC)$ for each $\ell \in \mathcal{L}$, we make the following assumptions:

A1 For each $\ell \in \mathcal{L}$ and a family $\mathcal{F}_i^\ell$, the error rate $E(C_i^\ell)$ is such that $C_i^\ell$ takes values in a countable set, and the multiple testing procedure used for testing this family can control $E(C_i^\ell)$ at any desired level $\alpha$ under the dependence among the $p$-values in $\mathcal{F}_i^\ell$.

A2 For each $\ell \in \mathcal{L}$ and a family $\mathcal{F}_i^\ell$, the $p$-values of the hypotheses in $\mathcal{F}_i^\ell$ are indepen-
TreeC: a general selection-adjusted hierarchical procedure

**Input:** The target levels for error rates for \( \mathbb{E}(sC^\ell), \ell \in \mathcal{L} : \{ q_0^{(i)}, i \in \mathcal{L} \} \)

For each \( \ell \in \mathcal{L} \) and family \( \mathcal{F}_I^\ell \), the \( \mathbb{E}(C_i^\ell) \)-controlling multiple testing procedure for testing family \( \mathcal{F}_I^\ell \)

For each \( \ell \notin \mathcal{L}, \ell \leq K \), where \( K = \max \{ \ell, \ell \in \mathcal{L} \} \), and family \( \mathcal{F}_I^\ell \), the selection rule defining the set \( S_i^\ell \)

The \( p \)-values for all the hypotheses at levels 1, \ldots, \( K \)

begin
    \( S^0 \leftarrow i_0 \)
    \( q_{i_0} = q_0^{(1)} \)
    for \( \ell = 1, \ldots, K \) do
        \( S^\ell \leftarrow \emptyset \)
    end

end

for \( \ell = 1, \ldots, K \) do
    while \( S^{\ell - 1} \neq \emptyset \) do
        for \( i \in S^{\ell - 1} \) do
            If \( \ell \in \mathcal{L} \), apply the \( \mathbb{E}(C_i^\ell) \)-controlling multiple testing procedure at level \( q_i \) on the \( p \)-values of family \( \mathcal{F}_I^\ell \)
            If \( \ell \notin \mathcal{L} \), apply the assigned selection rule for selecting the hypotheses in family \( \mathcal{F}_I^\ell \), defining the set \( S_i^\ell \)
            \( S^\ell \leftarrow S^\ell \cup S_i^\ell \)
        end
    end
end

Output: The set of all the selected hypotheses, \( \bigcup_{\ell=1}^K S^\ell \).

36
dent of the \( p \)-values of the hypotheses at levels \( 1, \ldots, \ell - 1 \) which are not ancestors of the family \( \mathcal{F}_\ell \).

A3 For each family \( \mathcal{F}_i^k \) where \( k \leq \max \{ \ell : \ell \in \mathcal{L} \} \) the selection rule defining the set \( S_i^k \) (which is defined by a multiple testing procedure for \( k \in \mathcal{L} \)) is a simple selection rule (see definition below).

As in [7] we note that for all practical purposes \( C_i^\ell \) is a count or a ratio of counts, therefore it takes values in a countable set, as required in A1. Below we accommodate to our setting the definition of a simple selection rule, given in [7]. In our context, selection of a parent hypothesis is equivalent to the selection of the family it indexes.

**Definition 1.** (simple selection rule). A selection rule is called simple if for each selected hypothesis \( H_j \in \mathcal{F}_i^\ell \), when the \( p \)-values belonging to \( \mathcal{F}_i^\ell \setminus \{ H_j \} \) are fixed, and the \( p \)-value of \( H_j \) can change as long as \( H_j \) is selected, the number of selected hypotheses in \( \mathcal{F}_i^\ell \), i.e. \( |S_i^\ell| \), remains unchanged.

The requirement that the selection rule is simple is a very lenient requirement. It is shown in [7] that any step-up or step-down non-adaptive multiple testing procedure defines a simple selection rule. Therefore the BH procedure used in the treeBH procedure defines a simple selection rule. When the selection of hypotheses is not made by multiple testing, when each hypothesis \( H_i \) is selected only based on the \( p \)-values in the family it indexes, \( \mathcal{F}_i^\ell \), the selection rule is simple.

**Theorem 3.** If assumptions A1–A3 hold, the TreeC procedure with input parameters \( q^{(i)}, i \in \mathcal{L} \), guarantees for each \( \ell \in \mathcal{L} \) that

\[ E(sC_i^\ell) \leq q^{(\ell)}. \]

**Proof of Theorem 3** The proof is similar to the proof in [7]. Let \( \ell \in \mathcal{L} \) be arbitrary fixed. Let \( C_i^+ \) be the countable support of \( C_i^\ell \) for each family \( \mathcal{F}_i^\ell \). Since the selection rule for selecting the hypotheses within each family is a simple selection rule, we can define the following event for each family \( \mathcal{F}_i^k \), where \( k \leq \max \{ \ell : \ell \in \mathcal{L} \} \), and for each \( H_j \in \mathcal{F}_i^k \), on the space of \( p \)-values for hypotheses in \( \mathcal{F}_i^k \setminus \{ H_j \} \): if \( H_j \) is selected, then \( |S_i^k| = r_i \), i.e. the number of hypotheses selected in \( \mathcal{F}_i^k \) is \( r_i \) including \( H_j \). Denote this event by \( C_i^{(j)} \).

\[
E(sC_i^\ell) = \\
\sum_{i_1 \in \mathcal{F}_1^1} \sum_{r_1 = 1}^{|\mathcal{F}_1^1|} \sum_{i_2 \in \mathcal{F}_1^2} \sum_{r_2 = 1}^{|\mathcal{F}_1^2|} \cdots \sum_{i_{\ell-1} \in \mathcal{F}_{i_{\ell-2}}^{\ell-2}} \sum_{r_{\ell-1} = 1}^{|\mathcal{F}_{i_{\ell-2}}^{\ell-2}|} \sum_{c \in C_i^{(j)}} \prod_{k=1}^{\ell-1} r_k \prod_{i \in \mathcal{L}} \prod_{k=1}^{\ell-1} r_k
\]

37
The TreeC procedure does not make any rejections in families that are not selected, therefore if the family $\mathcal{F}_{i_{\ell-1}}^\ell$ has an ancestor hypothesis that is not selected, $C_{i_{\ell-1}}^\ell = 0$. Therefore, for this procedure we obtain

$$
\mathbb{E}(sC^\ell) =
\sum_{i_1 \in F^1} \sum_{r_1 = 1}^{|F^1|} \sum_{i_2 \in F^2_{i_1}} \sum_{r_2 = 1}^{|F^2_{i_1}|} \cdots \sum_{i_{\ell-1} \in F_{i_{\ell-2}}^{\ell-1}} \sum_{r_{i_{\ell-1}} = 1}^{|F_{i_{\ell-2}}^{\ell-1}|}
\sum_{c \in C^\ell_{i_{\ell-1}}} c \mathbb{P}(C^\ell_{i_{\ell-1}} = c, \vert S_{i_k} \vert = r_k \text{ for } k = 1, \ldots, \ell - 1)
\prod_{k=1}^{\ell-1} r_k
= \sum_{c \in C^\ell_{i_{\ell-1}}} c \mathbb{P}(C^\ell_{i_{\ell-1}} = c, C_{r_k}^{(i_k)} \text{ for } k = 1, \ldots, \ell - 1)
\prod_{k=1}^{\ell-1} r_k
= \sum_{c \in C^\ell_{i_{\ell-1}}} c \mathbb{P}(C^\ell_{i_{\ell-1}} = c) \mathbb{P}(C^{(i_k)}_{r_k} \text{ for } k = 1, \ldots, \ell - 1)
\prod_{k=1}^{\ell-1} r_k
\mathbb{E}(C^\ell_{i_{\ell-1}}) \mathbb{P}(C^{(i_k)}_{r_k} \text{ for } k = 1, \ldots, \ell - 1)
\prod_{k=1}^{\ell-1} r_k
$$

(14)

The equality in (14) follows from the independence between the $p$-values in $F_i^\ell$ and the $p$-values of hypotheses at levels 1, $\ldots$, $\ell - 1$, which are not ancestors of $F_i^\ell$, for each family $\mathcal{F}_i^\ell$. In expression (15), $C^\ell_{i_{\ell-1}}$ is the value of the random measure of errors assigned to family $\mathcal{F}_{i_{\ell-1}}^\ell$ (where $\mathcal{F}_{i_0}^1 = F^1$) when a valid $\mathbb{E}(C^\ell_{i_{\ell-1}})$-controlling procedure at level $\ell$. Since $C^\ell_{i_{\ell-1}} = 0$ for families that are not selected, we obtain that $\mathbb{E}(C^\ell_{i_{\ell-1}}) \leq \frac{\prod_{k=1}^{\ell-1} r_k}{\prod_{k=1}^{\ell-1} |F_{i_{k-1}}^{\ell-1}|} q^{(\ell)}$ for each family $\mathcal{F}_i^\ell$. Using this inequality and the fact that for each sequence $i_1, i_2, \ldots, i_{\ell-1}$ such that $i_1 \in F^1$, $i_2 \in F^2_{i_1}$, $\ldots$, $i_{\ell-1} \in F_{i_{\ell-2}}^{\ell-1}$

$$
\sum_{r_1} \sum_{r_2} \cdots \sum_{r_{i_{\ell-1}} = 1}^{|F_{i_{\ell-2}}^{\ell-1}|}
\mathbb{P}(C^{(i_k)}_{r_k} \text{ for } k = 1, \ldots, \ell - 1) = 1
$$

38
we obtain

\[
\sum_{i_1 \in \mathcal{F}_1} \sum_{r_1 = 1}^{\mathcal{F}_1} \sum_{i_2 \in \mathcal{F}_2_{i_1}} \sum_{r_2 = 1}^{\mathcal{F}_2_{i_1}} \cdots \sum_{i_{\ell-1} \in \mathcal{F}_{\ell-2} r_{\ell-2} = 1}^{\mathcal{F}_{\ell-2} r_{\ell-2}} \sum_{r_{\ell-1} = 1}^{\mathcal{F}_{\ell-1}} \frac{\mathbb{E}(C_{i_{\ell-1}}^\ell) \mathbb{P}(C_{i_{\ell-1}k}^{i_k}) \text{ for } k = 1, \ldots, \ell - 1)}{\prod_{k=1}^{\ell-1} r_k} \leq \\
\sum_{i_1 \in \mathcal{F}_1} \sum_{i_2 \in \mathcal{F}_2_{i_1}} \cdots \sum_{i_{\ell-1} \in \mathcal{F}_{\ell-2} r_{\ell-2} = 1}^{\mathcal{F}_{\ell-2} r_{\ell-2}} \frac{\prod_{k=1}^{\ell-1} r_k}{\prod_{k=1}^{\ell-1} |\mathcal{F}_{k_{i_{k-1}}}^{i_{k-1}}|} q^{(\ell)} \frac{1}{\prod_{k=1}^{\ell-1} r_k} \sum_{r_{\ell-1} = 1}^{\mathcal{F}_{\ell-1}} \sum \cdots \sum \mathbb{P}(C_{i_{\ell-1}k}^{i_k}) \text{ for } k = 1, \ldots, \ell - 1) = \\
q^{(\ell)} \sum_{i_1 \in \mathcal{F}_1} \frac{1}{|\mathcal{F}_1|} \sum_{i_2 \in \mathcal{F}_2_{i_1}} \frac{1}{|\mathcal{F}_2_{i_1}|} \cdots \sum_{i_{\ell-1} \in \mathcal{F}_{\ell-2} r_{\ell-2} = 1}^{\mathcal{F}_{\ell-2} r_{\ell-2}} \frac{1}{|\mathcal{F}_{\ell-1}|} = q^{(\ell)}.
\]

Combining this result with the result in (15) we obtain \( \mathbb{E}(sC^\ell) \leq q^{(\ell)} \), thus the proof is complete. \( \Box \)
Supplementary figures

Figure S1: Pattern of nonzero coefficients across 50 causal SNPs (rows) and 250 simulated traits (columns) for a single tissue for the simulation described in Section 5.3. Simulated traits with no genetic basis are included in the illustration (i.e. blank columns), but non-causal SNPs (i.e. blank rows, which are in fact much more numerous than the causal SNPs) are omitted. Note that in each iteration of the simulation, the causal SNPs are selected at random, rather than being taken as a contiguous ordered block.