On Procedures Controlling the FDR for Testing Hierarchically Ordered Hypotheses

Gavin Lynch

Catchpoint Systems, Inc., 228 Park Ave S ♯28080

New York, NY 10003, U.S.A.

Wenge Guo*

Department of Mathematical Sciences

New Jersey Institute of Technology

Newark, NJ 07102, U.S.A.

Abstract

Complex large-scale studies, such as those related to microarray data and fMRI studies, often involve testing multiple hierarchically ordered hypotheses. However, most existing false discovery rate (FDR) controlling procedures do not exploit the inherent hierarchical structure among the tested hypotheses. In this paper, we first present a generalized stepwise procedure which generalizes the usual stepwise procedure to the case where each hypothesis is tested with a different set of critical constants. This procedure is helpful in creating a general framework under which our hierarchical testing procedures are developed. Then, we present several hierarchical testing procedures which control the FDR under various forms of dependence such as positive dependence and block dependence. Our simulation studies show that these proposed methods can be more powerful in some situations than alternative methods such as Yekutieli's hierarchical testing procedure (Yekutieli, *JASA* **103** (2008) 309-316). Finally, we apply our proposed procedures to a real data set involving abundances of microbes in different ecological environments.

Notation Index

The following summarizes commonly used notation and lists where each symbol is found.

Symbol	Description	Section	Page
$\mathcal{M}, m$	The set of tested hypotheses $\{H_1, \ldots, H_m\}$ and its cardinality.	2	5
$\mathcal{M}_i, m_i$	The set of descendant hypotheses of H_i and its cardinality.	2	5
$\mathcal{D}_i, d_i$	The set of ancestor hypotheses of H_i and its cardinality, also referred to as its depth.	2	5
$T(\cdot)$	A function that takes an index of a hypothesis and returns the index of its parent hypothesis.	2	5
$\mathcal{F}_d$	The set of hypotheses with depth d, $\mathcal{F}_d = \{H_i : d_i = d\}$.	2	5
$\mathcal{G}_d$	The union of $\mathcal{F}_1, \ldots, \mathcal{F}_d$.	A.2	25
D	The maximum depth of the hypotheses $\mathcal{G}_d \subseteq \mathcal{M}$ so that $\mathcal{G}_D = \mathcal{M}$.	2	4
ℓ	The total number of leaf hypotheses.	2	5
ℓ_i	The number of leaf hypotheses in set $\mathcal{M}_i$.	2	5
$R(\mathcal{A}), R$	The number of rejected hypotheses belonging to any set $\mathcal{A}$ and $R = R(\mathcal{M})$.	2	6
$V(\mathcal{A}), V$	The number of falsely rejected hypotheses belonging to any set $\mathcal{A}$ and $V = V(\mathcal{M})$.	2	6
$\alpha_i(\cdot)$	The critical function for testing the ith hypothesis H_i.	2	6

1. Introduction

In many problems involving the testing of multiple hypotheses, the hypotheses have an intrinsic, hierarchical structure such as a tree-like or graphical structure. These hierarchical structures often arise in multiple testing problems involving clinical trials (Mehrotra and Heyse, 2004; Dmitrienko et al., 2007; Huque and Alosh, 2008), genomics research (Yekutieli et al., 2006; Goeman and Mansmann, 2008; Heller et al., 2009; Guo et al., 2010) and fMRI studies (Benjamini and Heller, 2007). In general, hierarchical testing typically occurs while testing hierarchically structured hypotheses where, upon the rejection of one hypothesis, followup hypotheses are to be tested. For instance, Heller et al. (2009) introduced a hierarchical testing approach for analyzing microarray data where individual genes were grouped into gene sets. The gene sets were tested and upon successfully rejecting a gene set, the associated individual genes were tested. Guo et al. (2010) and Mehrotra and Heyse (2004) used a similar hierarchical testing approach for time-course microarray data and clinical safety data, respectively. Benjamini and Heller (2007) used a hierarchical testing approach to study fMRI data where the brain was divided into brain regions and each brain region was tested for significance. If a brain region was significant, the voxels within the brain region were tested. In addition, Meinshausen (2008) introduced a hierarchical testing method for addressing the problem of variable selection in multiple linear regression models.

In the field of multiple testing, the problem of controlling the familywise error rate (FWER) for testing hierarchically ordered hypotheses has received considerable attention (Dmitrienko et al., 2006, 2007; Goeman and Mansmann, 2008; Huque and Alosh, 2008; Meinshausen, 2008; Brechenmacher et al., 2011; Goeman and Finos, 2012); however, the FWER control can be too conservative for large-scale multiple testing. There has been very few work towards developing general methods for testing hierarchically ordered hypotheses that control the false discovery rate (FDR), even though the FDR is a more appropriate error measure for large scale multiple testing. To our knowledge, only Yekutieli (2008) has provided a general method for testing hierarchically ordered hypotheses that is specifically intended for controlling the FDR. Yekutieli's procedure, which is based on the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995), is only shown to control the FDR under independence. Some of the aforementioned procedures (Mehrotra and Heyse, 2004; Benjamini and Heller, 2007; Heller et al., 2009; Guo et al., 2010) can only be applied to special hierarchies consisting of only two layers.

In this paper, we propose new FDR controlling methods for testing hierarchically ordered hypotheses under various dependencies. Our approach towards controlling the FDR for testing hierarchically ordered hypotheses is different from that of Yekutieli's. First, to assist in the development of our hierarchical testing procedures, we introduce a new concept of generalized stepwise procedure, which generalizes the usual stepup, stepdown, and stepup-down procedures to the case where each hypothesis is tested with a different set of critical constants. The hypotheses are organized into different families according to their depth in the hierarchical structure. The formed families are sequentially tested by using the generalized stepwise procedures for which the corresponding critical constants take into account of the testing outcomes of higher-ranked families. Based on this approach, we were able to develop several new hierarchical testing procedures which control the FDR under various dependence structures including positive dependence and block dependence. To our knowledge, the procedures are the first procedures developed for testing hierarchically ordered hypotheses with proven control of the FDR under dependence structures other than independence. Furthermore, our simulation study shows that these procedure are quite powerful. The most powerful procedure, which we prove controls the FDR under positive block dependence, significantly outperforms Yekutieli's procedure in terms of power even though Yekutieli's procedure is only shown to control the FDR under independence, which is a special case of positive block dependence.

Another interesting finding of this research is that when the hierarchy takes on some special configurations, our procedures reduce to the existing FDR controlling procedures. For example, when there is no hierarchical structure, our proposed procedures reduce to the BH procedure and the Benjamini-Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001). When the hierarchy takes on a fixed sequence structure, our procedures are equivalent to the fixed sequence procedures in Lynch et al. (2016). This shows that our procedures are the combination of stepwise and fixed sequence methods.

The rest of this paper is outlined as follows. In Section 2, we provide relevant notation and definitions that will be used throughout this paper. Section 3 presents our proposed generalized stepwise procedure. Section 4 presents our new hierarchical testing procedures with proven control of the FDR under various dependencies. Sections 5 and 6 present a simulation study and real data analysis where we compare our procedures with Yekutieli's procedure. Finally, Section 7 provides some brief discussions.
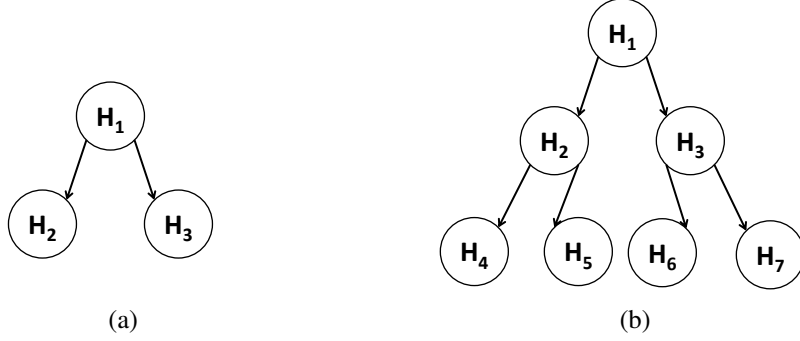
4

Figure 1: (a) An example of a hierarchical structure with 3 hypotheses for which H_2 and H_3 are only tested if H_1 is rejected. (b) An example of a hierarchical structure with 7 hypotheses for which H_2 and H_3 are only tested if H_1 is rejected, H_4 and H_5 are only tested if H_2 is rejected, and H_6 and H_7 are only tested if H_3 is rejected.

2. Preliminaries

Suppose there are m hypotheses $H_1, \ldots, H_m$ to be tested that are organized hierarchically in a tree-like structure where each hypothesis can have several child hypotheses but at most one parent hypothesis. Let $\mathcal{M} = \{H_1, \ldots, H_m\}$ be the set of the m tested hypotheses. Let $T : \{0, \ldots, m\} \to \{0, \ldots, m\}$ be a function that takes an index of a hypothesis and returns the index of the parent hypothesis with $T(0) = 0$. That is, if H_i has a parent hypothesis, its parent hypothesis is $H_{T(i)}$; otherwise H_i does not have a parent hypothesis and $T(i) = 0$. Define $T^0(i) = i$ and $T^k(i) = T(T^{k-1}(i))$ for any positive integer k. Let $\mathcal{D}_i = \{H_j : T^k(i) = j$ for $k = 0, \ldots, m\}$ so that $\mathcal{D}_i$ is the set of all ancestor hypotheses of H_i, which includes H_i. Let d_i be the cardinality of $\mathcal{D}_i$, $d_i = |\mathcal{D}_i|$. The depth of H_i in the hierarchy is defined as d_i. Let D be the maximum depth of the m hypotheses to be tested. If $d_i = 1$, then H_i does not have a parent hypothesis. Let $\mathcal{M}_i = \{H_j : T^k(j) = i$ for $k = 0, \ldots, m\}$ so that $\mathcal{M}_i$ is the set of all descendant hypotheses of H_i, which also includes H_i. We will refer to the hypotheses in set $\mathcal{M}_i$ as the subtree under H_i. Let m_i be the cardinality of $\mathcal{M}_i$, $m_i = |\mathcal{M}_i|$. If $m_i = 1$, then H_i has no children and it is referred to as a leaf hypothesis. We denote the number of leaf hypotheses in the whole hierarchy by ℓ and the number of leaf hypotheses in the subtree under H_i by ℓ_i. Formally, $\ell = \sum_{H_j \in \mathcal{M}} I\{m_j = 1\}$ and $\ell_i = \sum_{H_j \in \mathcal{M}_i} I\{m_j = 1\}$. Our procedures introduced in Section 4 group the hypotheses into D families by depth where family d contains all hypotheses with depth d, that is, $\mathcal{F}_d = \{H_i \in \mathcal{M} : d_i = d\}$. For example, in Figure 1(a), $T(2) = T(3) = 1$ and H_2 and H_3 are leaf hypotheses. In Figure 1(b), $T(6) = T(7) = 3, \mathcal{D}_6 = \{H_1, H_3, H_6\}, \mathcal{M}_2 = \{H_2, H_4, H_5\}$, and $\mathcal{F}_3 = \{H_4, H_5, H_6, H_7\}$.

The hypotheses in the hierarchical structure are tested hierarchically by a testing proce-

dure based on their corresponding p-values $P_1, \ldots, P_m$. By hierarchical testing, we mean a hypothesis is only tested if its parent hypothesis has been rejected or it does not have a parent hypothesis. For any set $\mathcal{A} \subseteq \mathcal{M}$, define $R(\mathcal{A})$ and $V(\mathcal{A})$ to be the number of rejected hypotheses and falsely rejected hypotheses among the hypotheses in set $\mathcal{A}$, respectively. For example, $R(\mathcal{M})$ and $V(\mathcal{M})$ are the number of rejected hypotheses and falsely rejected hypotheses among all the m tested hypotheses, respectively, and $R(\mathcal{M}_i)$ and $V(\mathcal{M}_i)$ are number of rejected hypotheses and falsely rejected hypotheses among the hypotheses in the subtree $\mathcal{M}_i$, respectively. For simplicity, often we will use R and V to denote $R(\mathcal{M})$ and $V(\mathcal{M})$, respectively. The FWER of this procedure is defined as $\Pr(V > 0)$. The FDR of this procedure is defined as FDR $= E(V/R)$, where we use the convention that $V/R = 0$ when $R = 0$. In addition, we will always use $|\mathcal{A}|$ to denote the cardinality of any set $\mathcal{A}$ throughout the paper.

Most existing multiple testing procedures are stepwise methods which are based on the ordered p-values $P_{(1)} \leq \cdots \leq P_{(m)}$ with corresponding hypotheses $H_{(1)}, \ldots, H_{(m)}$. Typically the rejection thresholds of a stepwise procedure are based on a sequence of non-decreasing critical constants but in this paper, for convenience, we will instead test the hypotheses using a non-decreasing, non-negative function $\alpha_0 : \{0, \ldots, m+1\} \to \mathbb{R}$ called a critical function where $\alpha_0(0) = 0$. For example, the critical function of the BH procedure is $\alpha_0(r) = r\alpha/m$. A stepwise procedure first determines the number of rejections R based on the critical function, then for each $i = 1, \ldots, m$, it rejects H_i if $P_i \leq \alpha_0(R)$ and accepts H_i if $P_i > \alpha_0(R)$. With $P_{(0)} \equiv 0$ and $P_{(m+1)} \equiv \infty$, a stepup procedure sets $R = \max\{0 \leq r \leq m : P_{(r)} \leq \alpha_0(r)\}$. A stepdown procedure sets $R = \min\{1 \leq r \leq m+1 : P_{(r)} > \alpha_0(r)\} - 1$. Finally, a stepup-down procedure of order k, which generalizes stepup and stepdown procedures, sets $R = \max\{0 \leq r \leq k-1 : P_{(r)} \leq \alpha_0(r)\}$ if $P_{(k)} > \alpha_0(k)$ and $R = \min\{k+1 \leq r \leq m+1 : P_{(r)} > \alpha_0(r)\} - 1$ if $P_{(k)} \leq \alpha_0(k)$. When $k = m$, the stepup-down procedure reduces to the stepup proceudre and when $k = 1$, it reduces to the stepdown procedure. It should be noted that the event $\{P_{(r)} \leq \alpha_0(r)\}$ is equivalent to the event $\{r \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_0(r)\}\}$. Thus, the number of rejections can also be expressed by

$$R = \max\left\{0 \leq r \leq m : r \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_0(r)\}\right\} \tag{1}$$

for the stepup procedure,

$$R = \min \left\{ 1 \le r \le m+1 : r > \sum_{i=1}^{m} I\{P_i \le \alpha_0(r)\} \right\} - 1 \qquad (2)$$

for the stepdown procedure, and

$$R = \begin{cases} \max\left\{0 \le r \le k-1 : r \le \sum_{i=1}^{m} I\{P_i \le \alpha_0(r)\}\right\} & \text{if } k > \sum_{i=1}^{m} I\{P_i \le \alpha_0(k)\} \\ \min\left\{k+1 \le r \le m+1 : r > \sum_{i=1}^{m} I\{P_i \le \alpha_0(r)\}\right\} - 1 & \text{if } k \le \sum_{i=1}^{m} I\{P_i \le \alpha_0(k)\} \end{cases} \qquad (3)$$

for the stepup-down procedure of order k. Refer to Tamhane et al. (1998) and Sarkar (2002) for further discussion on stepup-down procedure.

Throughout this paper we make use of the following basic assumption regarding marginal p-values: for any p-value P_i with H_i being true,

$$\Pr\left(P_i \le p\right) \le p \quad \text{for any } 0 \le p \le 1. \qquad (4)$$

We consider several types of joint dependence throughout this paper: arbitrary dependence, positive dependence, and block dependence. Under arbitrary dependence, the p-values are not known to have any specific type of dependence structure. Positive dependence and block dependence are characterized by the following assumptions.

Assumption 1. Positive Dependence Assumption
For any coordinatewise non-decreasing function of the p-values ψ,

$$E\left(\psi(P_1, \ldots, P_m) \mid P_i \le p\right) \text{ is non-decreasing in } p \text{ for each } p\text{-value } P_i \text{ such that } H_i \text{ is true.}$$

$$(5)$$

Assumption 2. Block Dependence Assumption
For each $d = 1, \ldots, D$, the p-values corresponding to the hypotheses in $\mathcal{F}_d$ are independent of the p-values corresponding to the hypotheses not in $\mathcal{F}_d$.

Assumption 1 is slightly more relaxed than the condition of positive regression dependence on a subset (PRDS) introduced in Benjamini and Yekutieli (2001). Assumption 2 only characterizes the joint dependence of the p-values across families but does not describe the joint dependence within families.

3. Generalized Stepwise Procedure

In order to present our hierarchical testing procedures in the next section, in this section, we present a new type of procedure termed as generalized stepwise procedure, including generalized stepup, stepdown, and stepup-down procedures, which generalizes the usual stepup, stepdown, and stepup-down procedures. In a non-hierarchical multiple testing problem where a stepwise procedure is used to test the hypotheses, the tested hypotheses often have the same importance and thus, it is natural to test those hypotheses with the same critical function, as shown in (3). However, when the hypotheses have a hierarchical structure, the importance of a hypothesis depends on where it is located in the hierarchy. Hence, for a desired procedure, each hypothesis should be tested with a different critical function that reflects its importance, and so we generalize the usual stepwise procedure as follows.

Given m non-decreasing critical functions $\alpha_i(r), i = 1, \ldots, m$, our proposed generalized stepwise procedure rejects H_i if $P_i \leq \alpha_i(R)$ for each $i = 1, \ldots, m$ where R is determined as follows. For the generalized stepup procedure,

$$R = \max\left\{0 \leq r \leq m : r \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\}\right\}, \tag{6}$$

for the generalized stepdown procedure,

$$R = \min\left\{1 \leq r \leq m+1 : r > \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\}\right\} - 1, \tag{7}$$

and for the generalized stepup-down procedure of order k,

$$R = \begin{cases} \max\left\{0 \leq r \leq k-1 : r \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\}\right\} & \text{if } k > \sum_{i=1}^{m} I\{P_i \leq \alpha_i(k)\} \\ \min\left\{k+1 \leq r \leq m+1 : r > \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\}\right\} - 1 & \text{if } k \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_i(k)\}. \end{cases} \tag{8}$$

It is easy to see that when $\alpha_i(r) = \alpha_0(r)$ for each $i = 1, \ldots, m$, (6), (7), and (8) reduce to (1), (2), and (3), respectively. Thus, the generalized stepup, stepdown, and stepup-down procedures reduce to the usual stepwise procedures, respectively. It should be noted that when $k = m$, (8) reduces to (6) and when $k = 1$, (8) reduces to (7).

The generalized stepwise procedure is fairly general and we present two examples to show its broad applicability.

Example 1. Consider a weighted multiple testing problem where H_i has corresponding weight $w_i, i = 1, \ldots, m$. A weighted stepwise procedure with the critical function $\alpha_0(r)$ tests H_i based on weight-adjusted p-values P_i/w_i instead of P_i. This is equivalent to a generalized stepwise procedure with the critical functions $\alpha_i(r) = w_i\alpha_0(r), i = 1, \ldots, m$ so that the weighted stepwise procedure can be regarded as a special case of the generalized stepwise procedure.

Example 2. Fixed sequence procedures assume the testing order of the hypotheses has been specified a-priori and that H_i is not tested unless $H_1, \ldots, H_{i-1}$ have all been rejected. Lynch et al. (2016) showed that the fixed sequence procedure that rejects H_i when $P_i \leq m\alpha/(m-i+1)$ controls the FDR at level α under arbitrary dependence. This procedure is a special case of the generalized stepdown procedure with critical functions $\alpha_i(r) = I\{r \geq i\}m\alpha/(m-r+1)$. Other fixed sequence procedures can be defined similarly.

From (8), it can be seen that many of the familiar properties of stepwise procedures also hold for the generalized stepwise procedure. For example, the number of rejections R is a coordinatewise non-increasing function of the p-values and R is a non-decreasing function of k (i.e. a stepup-down procedure of order k rejects more hypotheses than a stepup-down procedure of order $k-1$). The most important property is a self-consistency property which allows us to express R as

$$R = \sum_{i=1}^{m} I\{P_i \leq \alpha_i(R)\}. \tag{9}$$

(Blanchard and Roquain (2008) discussed a weaker self-consistency condition for the usual stepwise procedure with the critical function $\alpha_0(r)$, which is the inequality $R \leq \sum_{i=1}^{m} I\{P_i \leq \alpha_0(R)\}$). This property ensures that R as determined in (8) is indeed the number of rejections by the generalized stepwise procedure. Thus, the event $\{H_i \text{ is rejected}\}$ can be expressed as $\{P_i \leq \alpha_0(R)\}$ with R being the number of rejections. To see why this property holds, let us define $\psi(r) = \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\}$. When $k > \psi(k)$, then $R = \max\{0 \leq r \leq k-1 : r \leq \psi(r)\}$ and if $k \leq \psi(k)$, then $R + 1 = \min\{k + 1 \leq r \leq m+1 : r > \psi(r)\}$. In either case, it is easy to see that $R \leq \psi(R)$ and $R + 1 > \psi(R+1)$. The fact that $\psi(R+1) < R+1$ implies $\psi(R+1) \leq R$. Thus, $R = \psi(R)$ since $R \leq \psi(R) \leq \psi(R+1) \leq R$.

To conclude this section, we present an efficient algorithm for finding the number of rejections by the generalized stepwise procedure. The algorithm is particularly useful when the number of hypotheses is very large.

Algorithm 1. Given a positive integer $1 \leq k \leq m$ and critical functions $\alpha_i(\cdot), i = 1, \ldots, m$,

define $\psi(r) = \sum_{i=1}^{m} I\{P_i \leq \alpha_i(r)\}$.

1. Let $t = 1$ and $r_t = k$.

2. If $r_t > \psi(r_t)$, then

    (a) Increase t by 1 and set $r_t = \psi(r_{t-1})$.

    (b) If $r_t \leq \psi(r_t)$, then let $R = r_t$ and stop; otherwise, if $r_t > \psi(r_t)$, repeat step 2(a).

3. Otherwise, if $r_t \leq \psi(r_t)$, then

    (a) Increase t by 1 and set $r_t = \psi(r_{t-1}) + 1$.

    (b) If $r_t > \psi(r_t)$, then let $R = r_t - 1$ and stop; otherwise, if $r_t \leq \psi(r_t)$, repeat step 3(a).

Proposition 1. *The value of R in (8) can be solved by algorithm 1.*

Proof. The proof is in the Appendix. ∎

4. Hierarchical FDR Control

In this section, we describe our procedure to test hierarchically ordered hypotheses. The tested hypotheses are arranged into D families, $\mathcal{F}_1, \ldots, \mathcal{F}_D$, where $\mathcal{F}_d$ is the family of hypotheses with depth d. Given m non-decreasing critical functions $\alpha_i(r), i = 1, \ldots, m$, the hypotheses are tested as follows.

Definition 1. General Hierarchical Testing Procedure

1. *Test $\mathcal{F}_1$ by using the generalized stepup procedure with critical functions $\alpha_i(r), H_i \in \mathcal{F}_1$. Let $\mathcal{S}_1$ be the set of rejected hypotheses and $R(\mathcal{F}_1)$ be the number of rejected hypotheses in $\mathcal{F}_1$. Test $\mathcal{F}_2$.*

2. *To test $\mathcal{F}_d$, use the generalized stepup procedure with critical functions*

$$\alpha_i^*(r) = I\{H_{T(i)} \text{ is rejected}\}\alpha_i \left(r + \sum_{j=1}^{d-1} R(\mathcal{F}_j) \right), \quad H_i \in \mathcal{F}_d.$$

Let $\mathcal{S}_d$ be the set of rejected hypotheses and $R(\mathcal{F}_d)$ be the number of rejected hypotheses in $\mathcal{F}_d$. Test $\mathcal{F}_{d+1}$.

3. *The set of rejected hypotheses is $\bigcup_{d=1}^{D} \mathcal{S}_d$ and the total number of rejections is $R = \sum_{d=1}^{D} R(\mathcal{F}_d)$.*

The above procedure is termed as a hierarchical testing procedure since the procedure will accept any hypothesis whose parent hypothesis has been accepted. This can be seen in the construction of the critical functions in step 2 where $\alpha_i^*(r) = 0$ if H_i's parent hypothesis $H_{T(i)}$ has not been rejected so that H_i cannot be rejected. It should be noted that the parents of all hypotheses in $\mathcal{F}_d$ are in $\mathcal{F}_{d-1}$, which is tested before testing $\mathcal{F}_d$. Hence, for each $H_i \in \mathcal{F}_d$, the event $\{H_{T(i)}$ is rejected$\}$ is observed by the time $\mathcal{F}_d$ is tested.

Remark 1. In Definition 1, when all the hypotheses in $\mathcal{F}_d$ have the same critical functions and every $H_i \in \mathcal{F}_d$ can be tested (i.e. $H_{T(i)}$ is rejected), the generalized stepup procedure used for testing $\mathcal{F}_d$ reduces to the usual stepup procedure. However, our critical functions for testing hierarchically ordered hypotheses, which are presented in the next subsections, are not the same and depend on where the hypothesis is located in the hierarchy. Furthermore, since the hypotheses in $\mathcal{F}_d$ may not have the same parent, the parent hypotheses $H_{T(i)}$ could be rejected for some, but not all, of $H_i \in \mathcal{F}_d$. Hence, only in an uncommon case does the generalized stepup procedure reduce to the usual stepup procedure for testing $\mathcal{F}_d$.

Following (9), the hierarchical testing procedure has the following self-consistency property in each family $\mathcal{F}_d$,

$$R(\mathcal{F}_d) = \sum_{H_i \in \mathcal{F}_d} I\{P_i \leq \alpha_i^*(R(\mathcal{F}_d))\}, d = 1, \ldots, D,$$

where $\alpha_i^*(r) = I\{H_{T(i)}$ is rejected$\}\alpha_i(r + \sum_{j=1}^{d-1} R(\mathcal{F}_j))$. Hence, the event $\{H_i$ is rejected$\}$ is equivalent to the event $\{H_{T(i)}$ is rejected$, P_i \leq \alpha_i(\sum_{j=1}^{d_i} R(\mathcal{F}_j))\}$, where $\sum_{j=1}^{d_i} R(\mathcal{F}_j)$ is the number of rejections in the first d_i families, $\mathcal{F}_1, \ldots, \mathcal{F}_{d_i}$. This property will be useful to prove the FDR control of our procedures.

Now that we have defined our hierarchical testing procedure, we will consider various dependence structures, such as positive dependence, arbitrary dependence, and block dependence, and develop newer hierarchical testing procedures which control the FDR under these dependence structures. The proofs of all the theorems in this section are in the appendix.

4.1. Procedure under Positive Dependence. We first consider positive dependence structure. Positive dependence has received much attention in multiple testing due to the fact that

several popular multiple testing procedures have been developed under this type of dependence (see Sarkar (1998); Benjamini and Yekutieli (2001); Sarkar (2002); Guo and Sarkar (2016)). Our procedure under positive dependence is as follows.

Theorem 1. *FDR Control under Positive Dependence*

Under Assumption 1, the hierarchical testing procedure with critical functions

$$\alpha_i(r) = \frac{\ell_i \alpha}{\ell} \frac{m_i + r - 1}{m_i}, \quad i = 1, \ldots, m,$$

strongly controls the FDR at level α.

Consider the special case when there is no hierarchical ordering (i.e. $\ell = m$ and $\ell_i = m_i = 1$) so that the hypotheses do not have any pre-defined structure and the problem reduces to a non-hierarchical multiple testing problem. We will refer to this configuration as the non-hierarchical configuration. Under this configuration, all the hypotheses belong to the same family, $\mathcal{F}_1$, so that the hierarchical testing procedure reduces to the usual stepup procedure which can be further reduced to a normal stepup procedure since all the critical functions are equal to $r\alpha/m$. Hence, the procedure reduces to the stepup procedure with critical function $r\alpha/m$, which is the BH procedure. Thus, our result generalizes the BH to the testing of hierarchically ordered hypotheses.

Now, we consider another special case where each family $\mathcal{F}_i$ has exactly one hypothesis $H_i, i = 1, \ldots, m$. Thus, the tested hypotheses $H_1, \ldots, H_m$ are pre-ordered, $\ell = 1$, and $m_i = m - i + 1$. We will refer to this configuration as the fixed sequence configuration. Under this configuration, the hierarchical testing procedure reduces to the fixed sequence method introduced in Lynch et al. (2016), where hypothesis H_i is rejected if, and only if, hypotheses $H_1, \ldots, H_{i-1}$ have all been rejected and $P_i \leq m\alpha/(m-i+1)$. This method is shown to control the FDR at level α under arbitrary dependence. Thus, our result also generalizes the fixed sequence procedure to the testing of hierarchically ordered hypotheses.

Remarkably, our result has connected two opposing testing methods: the testing of non-ordered hypotheses (through the BH procedure) and the testing of fully ordered hypotheses (through the fixed sequence procedure).

Finally, we consider a third configuration which we call the binary tree configuration. This configuration is helpful for evaluating the critical functions in the hierarchical setting and it is defined as follows. There is one hypothesis in $\mathcal{F}_1$ and each hypothesis has two child hypotheses except for the leaf hypotheses in $\mathcal{F}_D$. Hence, $\ell = 2^{D-1}$ and $m = 2^D - 1$. For

Table 1: A comparison of critical functions for the procedure in Theorem 1 and Meinshausen's procedure when testing the hypotheses in Figure 1(b).

		Theorem 1							Meinshausen
		$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	
Family 1	$\alpha_i(r)$	α	-	-	-	-	-	-	α
Family 2	$\alpha_i(r)$	-	$2\alpha/3$	$5\alpha/6$	-	-	-	-	$\alpha/2$
Family 3	$\alpha_i(r)$	-	-	$3\alpha/4$	α	$5\alpha/4$	$3\alpha/2$	$7\alpha/4$	$\alpha/4$

each $d = 1, \ldots, D$, there are 2^{d-1} hypotheses in $\mathcal{F}_d$ and for each $H_i \in \mathcal{F}_d$, $\ell_i = 2^{D-d}$ and $m_i = 2^{D-d+1} - 1$. Under this configuration, the critical functions of Theorem 1 are, after simplification,

$$\alpha_i(r) = \frac{\alpha}{2^{d-1}} \left(1 + \frac{r-1}{2^{D-d+1} - 1} \right), \quad H_i \in \mathcal{F}_d, d = 1, \ldots, D. \tag{10}$$

Compared to Meinshausen's FWER controlling hierarchical testing procedure, which is equivalent to the hierarchical testing procedure with critical functions $\ell_i \alpha/\ell, i = 1, \ldots, m$, the critical functions of Theorem 1 are $(m_i + r - 1)/m_i$ times larger for H_i. In the binary tree configuration, Meinshausen's critical function for H_i in $\mathcal{F}_d$ is $\alpha_i(r) = \alpha/2^{d-1}$, which is $(1 + (r-1)/(2^{D-d+1} - 1))$ times smaller than the critical function in Theorem 1. Table 1 lists the critical functions of Theorem 1 and Meinshausen's procedure for testing the hypotheses in Figure 1(b) which has the binary tree configuration. For family d, only the values of r between d and $\sum_{j=1}^{d} |\mathcal{F}_j|$ are listed in Table 1 due to the fact that if a hypothesis in family d is rejected, then all d of its ancestor hypotheses including the hypothesis itself are rejected so that $d \leq \sum_{j=1}^{d} R(\mathcal{F}_j) \leq \sum_{j=1}^{d} |\mathcal{F}_j|$.

Remark 2. It should be noted that the hierarchical testing procedure relies on the generalized stepup procedure to test each family; however, our proof of the FDR control for the procedure in Theorem 1 (and our proof of FDR control for the remaining procedures in this section) still holds if the generalized stepup-down procedure of any arbitrary order (including the generalized stepdown procedure) is used to test each family. Nevertheless, in practice we are generally trying to maximize the number of rejections subject to the FDR control. Since with the same critical functions, the generalized stepup procedure is more powerful than the corresponding generalized stepup-down and generalized stepdown procedures, we opted to use the generalized stepup procedure to test each family.

4.2. Procedure under Arbitrary Dependence. In this subsection we introduce a FDR controlling hierarchical testing procedure under arbitrary dependence. Since arbitrary dependence is a more general type of joint dependence than positive dependence, it follows that the procedure under arbitrary dependence will not be quite as powerful as the procedure from Theorem 1.

Theorem 2. *FDR Control under Arbitrary Dependence*
The hierarchical testing procedure with critical functions

$$\alpha_i(r) = \frac{\ell_i \alpha}{\ell} \frac{m_i + r - 1}{m_i} \frac{1}{c_i}, \quad where \ c_i = 1 + \sum_{j=d_i}^{|\mathcal{G}_{d_i}|-1} 1/(m_i + j),$$

for $i = 1, \ldots, m$, strongly controls the FDR at level α under arbitrary dependence.

Just like Theorem 1, we consider the non-hierarchical configuration of hypotheses. In this special case, all of the critical functions are $r\alpha/(mc)$ where $c = \sum_{j=1}^{m} 1/j$ so that this procedure reduces to the stepup procedure with critical function $r\alpha/(mc)$, which is the BY procedure. Thus, this result extends the BY procedure to the testing of hierarchically ordered hypotheses. On the other hand, we also consider the fixed sequence configuration. Here, the rejection threshold for H_i is $m\alpha/(m-i+1)$, which is the same as the procedure from Theorem 1 under this configuration.

It is easy to see that the critical functions of this procedure are scaled down compared with the procedure from Theorem 1 in order to ensure the FDR control under arbitrary dependence, similar to the way the BY procedure is scaled down compared with the BH procedure. Consider the example in Figure 1 (b) which consists of 7 hypotheses. Here, $c_1 = 1, c_2 = c_3 = 1.2$, and $c_4 = c_5 = c_6 = c_7 = 1.76$ which means the critical functions of Theorem 1 are as large, 1.2 times larger, and 1.76 times larger than the critical functions of Theorem 2 for testing $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$, respectively. The critical function of the BH procedure, on the other hand, is $\sum_{i=1}^{7} 1/i = 2.59$ times larger than the critical function of the BY procedure for testing 7 hypotheses in the non-hierarchical setting. This holds in general, that the constants in the critical functions of Theorem 2 are much smaller in the hierarchical setting than in the non-hierarchical setting (i.e. the BY procedure). It shows that the FDR controlling procedure under arbitrary dependence tends to be less affected by not having the assumption of positive dependence in the hierarchical setting than in the non-hierarchical setting.

4.3. Procedures under Block Dependence. In this subsection, we consider block dependence and develop more powerful versions of the procedures in Theorems 1 and 2 by taking this dependence into account. Since block dependence only describes the dependence of the p-values across families, we consider both positive dependence and arbitrary dependence to describe the dependence of the p-values within the families which we will refer to as block positive dependence and block arbitrary dependence, respectively.

In the fixed sequence configuration, block dependence reduces to independence. Under this configuration of hypotheses, both of our procedures presented in this subsection reduce to the more powerful FDR controlling fixed sequence procedure under independence, whereas the procedures in the last two subsections reduce to the less powerful FDR controlling fixed sequence procedure under arbitrary dependence (Lynch et al., 2016).

First, we consider block positive dependence.

Theorem 3. *FDR Control under Block Positive Dependence*

Under Assumption 1 and 2, the hierarchical testing procedure with critical functions

$$\alpha_i(r) = \begin{cases} \dfrac{\ell_i r \alpha}{\ell + \ell_i (r-1)\alpha} & \text{if } H_i \text{ is not a leaf hypothesis} \\ \dfrac{r\alpha}{\ell} & \text{if } H_i \text{ is a leaf hypothesis} \end{cases}$$

for $i = 1, \ldots, m$, strongly controls the FDR at level α.

In the non-hierarchical configuration, this procedure reduces to the BH procedure since all the critical functions are $r\alpha/m$. It should be noted that under this configuration there is only one family so that block dependence is irrelevant and we are left with just the positive dependence assumption. Thus, both this procedure and the procedure from Theorem 1, which both assume positive dependence, reduce to the BH procedure in the non-hierarchical configuration.

In the hierarchical setting, this procedure offers a large improvement over the critical functions of Theorem 1. To see this, consider the binary tree configuration. In this case, the critical functions are

$$\alpha_i(r) = \frac{r\alpha}{2^{d-1} + (r-1)\alpha}, \ H_i \in \mathcal{F}_d, d = 1, \ldots, D-1 \text{ and } \alpha_i(r) = \frac{r\alpha}{2^{D-1}}, \ H_i \in \mathcal{F}_D. \quad (11)$$

Comparing (10) to (11), one can see that (11) is, in general, much larger than (10), approximately $r/(1 + (r-1)/(2^{D-d+1} - 1))$ times larger when α is small. For example, when

15

$D = 5, d = 3$, and $r = 4$ the increase is by a factor of almost 3. Also, compared to Meinshausen's FWER controlling hierarchical testing procedure, which uses critical function $\ell_i \alpha / \ell, i = 1, \ldots, m$, the critical functions of Theorem 3 are approximately r times larger under every configuration for small α. Hence, the procedure from Theorem 3, which requires the strongest dependence assumption to control the FDR, is our most powerful hierarchical testing procedure.

Finally, we consider block arbitrary dependence.

Theorem 4. *FDR Control under Block Arbitrary Dependence*

Under Assumption 2, the hierarchical testing procedure with critical functions

$$
\alpha_i(r) = \begin{cases} \dfrac{\ell_i r \alpha}{\ell + \ell_i(r-1)\alpha} \dfrac{1}{c_i} & \text{if } H_i \text{ is not a leaf hypothesis} \\[3ex] \dfrac{r\alpha}{\ell} \dfrac{1}{c_i} & \text{if } H_i \text{ is a leaf hypothesis} \end{cases}
$$

where

$$
c_i = \begin{cases} 1 + \sum_{j=1}^{|\mathcal{F}_{d_i}|-1} \dfrac{\ell - \ell_i \alpha}{(j + d_i)(\ell + \ell_i(j + d_i - 2)\alpha)} & \text{if } H_i \text{ is not a leaf hypothesis} \\[3ex] 1 + \sum_{j=1}^{|\mathcal{F}_{d_i}|-1} \dfrac{1}{j + d_i} & \text{if } H_i \text{ is a leaf hypothesis} \end{cases}
$$

and $|\mathcal{F}_{d_i}|$ is the cardinality of $\mathcal{F}_{d_i}$, strongly controls the FDR at level α.

This procedure reduces to the BY procedure in the non-hierarchical configuration so that both procedures under arbitrary dependence reduce to the BY procedure. Similar to the procedures from Theorems 1 and 2, the critical functions of this procedure are a factor smaller than the critical functions of Theorem 3. Again, we consider the hypotheses in Figure 1 (b). Here, $c_1 = 1, c_2 = c_3 = 1.317$, and $c_4 = c_5 = c_6 = c_7 = 1.760$ at $\alpha = 0.05$. In this example, the c_i's are significantly smaller than the constant for the BY procedure for testing 7 hypotheses, which is 2.59. However, c_2 and c_3 for Theorem 4 are larger than c_2 and c_3 for Theorem 2, which are both 1.2, but this is not true in general for the c_i's. The portion of the critical function without c_i, is generally much larger for Theorem 4 than for Theorem 2 so that the procedure from Theorem 4 is typically more powerful than the procedure from Theorem 2.

Remark 3. Our proofs of the theorems in this section heavily rely on mathematical induction. The hierarchical structure of the hypotheses implies a recursive property where the hypotheses

Table 2: The procedure from Theorem 1 at level $\alpha = 0.05$ to hierarchically test the hypotheses presented in Figure 1(b) with p-values $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$.

Procedure 3		$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	Outcome
Family 1									
generalized stepup	$\alpha_i^*(R)$	0.05	-	-	-	-	-	-	$R=1$
Reject H_1 and set $R(\mathcal{F}_1) = 1$									
Family 2									
generalized stepup	$\alpha_i^*(R)$	-	0.033	0.033	-	-	-	-	$R=1$
Accept H_2, reject H_3 and set $R(\mathcal{F}_2) = 1$									
Family 3									
generalized stepup	$\alpha_i^*(R)$	-	-	-	0	0	0.05	0.05	$R=2$
Accept H_4 and H_5 and reject H_6 and H_7. Set $R(\mathcal{F}_3) = 2$									

in the subtree under any hypothesis also form a hierarchical structure. Hence, mathematical induction is a natural choice for proving results for hierarchical structures.

Below, we demonstrate how the hierarchical testing procedure in Theorem 1 works through an example as well as Yekutieli's and Meinshausen's hierarchical testing procedures.

Example 3. Consider the example presented in Figure 1(b). The maximum depth of the tree is 3 and the seven hypotheses in the tree are grouped as 3 families, which are $\{H_1\}$, $\{H_2, H_3\}$, and $\{H_4, H_5, H_6, H_7\}$. Suppose the p-values are $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$ and the hypotheses are tested using the procedure from Theorem 1, Yekutieli's procedure, and Meinshausen's procedure at level $\alpha = 0.05$.

Table 2 shows the value of the variables step-by-step for the procedure from Theorem 1. The first family is tested using the generalized stepup procedure and H_1, the only hypothesis in this family, is rejected. Now, $R(\mathcal{F}_1) = 1$. The second family is tested using the generalized stepup procedure with critical functions $\alpha_2^*(r) = \alpha_2(r + 1)$ and $\alpha_3^*(r) = \alpha_3(r + 1)$. H_3 can be rejected but H_2 cannot. Thus, $R(\mathcal{F}_2) = 1$. Finally, the third family is tested. Since H_2 was accepted and H_3 was rejected, we have $\alpha_4^*(r) = \alpha_5^*(r) = 0$, $\alpha_6^*(r) = \alpha_6(r + 2)$, and $\alpha_7^*(r) = \alpha_7(r + 2)$. Hypotheses H_6 and H_7 are rejected by the generalized stepup procedure.

Yekutieli's hierarchical testing procedure groups the hypotheses into families that share the same parent hypothesis so that the 4 families are $\{H_1\}$, $\{H_2, H_3\}$, $\{H_4, H_5\}$, and $\{H_6, H_7\}$. This procedure rejects hypotheses H_1 and H_3 (Table 3). Meinshausen's hierarchical testing procedure uses a fixed rejection threshold $\ell_i \alpha / \ell$ for testing H_i. This procedure rejects H_1 and

Table 3: Yekutieli's procedure at level $\alpha = 0.05$ to hierarchically test the hypotheses presented in Figure 1(b) with p-values $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03,$ and $p_7 = 0.05$.

Yekutieli's Procedure		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	Outcome
Family 1									
BH procedure	$\alpha_i(R)$	0.0174	-	-	-	-	-	-	$R = 1$
Reject H_1									
Family 2									
BH procedure	$\alpha_i(R)$	-	0.009	0.009	-	-	-	-	$R = 1$
Accept H_2 and reject H_3									
Family 3									
Not Tested		-	-	-	-	-	-	-	
Accept H_4 and H_5									
Family 4									
BH procedure	$\alpha_i(R)$	-	-	-	-	-	0	0	$R = 0$
Accept H_6 and H_7									

H_3 (Table 4).

5. Simulation Study

We conducted a simulation study to evaluate the performance of the proposed procedures. Specifically, the simulation study compared the performance of our proposed procedures, which are labeled Procedures 1-4 corresponding to the procedures introduced in Theorems 1-4, against Yekutieli's FDR controlling procedure in terms of the FDR control and average power. Several dependence configurations were considered as well as different hierarchical structures.

We generated m normal random variables with covariance matrix Σ and mean vector $\vec{\mu} = (\mu_1, \ldots, \mu_m)$ to test the m hypotheses $H_i : \mu_i \leq 0$ versus $H_i' : \mu_i > 0, i = 1, \ldots, m$. When H_i was true, we set $\mu_i = 0$. When H_i was false, we set μ_i to a positive value which was non-increasing in its depth d_i. Our intention was to simulate the setting where hypotheses that are near the top of the hierarchy are easier to reject than hypotheses near the bottom. As for the joint dependence, we considered a common correlation structure where Σ had off-diagonal components equal to ρ and diagonal components equal to 1. The p-value for testing the ith

Table 4: Meinshausen's procedure at level $\alpha = 0.05$ to hierarchically test the hypotheses presented in Figure 1(b) with p-values $p_1 = 0.01, p_2 = 0.75, p_3 = 0.008, p_4 = 0.6, p_5 = 0.85, p_6 = 0.03$, and $p_7 = 0.05$.

Meinshausen's Procedure		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	Outcome
Family 1									
Single Step	α_i	0.05	-	-	-	-	-	-	$R = 1$
Reject H_1									
Family 2									
Single Step	α_i	-	0.025	0.025	-	-	-	-	$R = 1$
Accept H_2 and reject H_3									
Family 3									
Single Step	α_i	-	-	-	-	-	0.0125	0.0125	$R = 0$
Accept H_4, H_5, H_6, and H_7									

leaf hypothesis was calculated using a one sided, one-sample Z-test.

We constructed two types of hierarchies: a shallow hierarchy and a deep hierarchy. Both hierarchies had 1000 leaf hypotheses.

The leaf hypotheses were randomly chosen with probability π_0 to be true and $1 - \pi_0$ to be false. Each non-leaf hypothesis was set to true only if all of its child hypotheses were true; otherwise it was set to false. For both hierarchies, the tree was balanced so that each parent hypothesis had the same number of child hypotheses. The two hierarchies are described in detail below.

Shallow Hierarchy: The maximum depth of this tree is 2 so that a hypothesis is either a leaf hypothesis or a top-level hypothesis with no parent. There are 10 top-level hypotheses each of which have 100 child hypotheses giving a total of 1010 hypotheses. For each false hypothesis H_i, $\mu_i = 3$ if $d_i = 1$ and $\mu_i = 2$ if $d_i = 2$.

Deep Hierarchy: The maximum depth of this tree is 4 and there are 8 top-level parents. Each parent hypothesis has 5 child hypotheses giving a total of 1248 hypotheses. For each false hypothesis H_i, $\mu_i = 3.5$ if $d_i = 1$, $\mu_i = 3$ if $d_i = 2$ or 3, and $\mu_i = 2$ if $d_i = 4$.

We set $\alpha = 0.05$ and for each procedure, we noted the false discovery proportion, which is the proportion of falsely rejected hypotheses among all rejected hypotheses, and the the proportion of rejected false null hypotheses among all false null hypotheses. Each tree was generated and tested 5000 times and the simulated values of the FDR and average power were
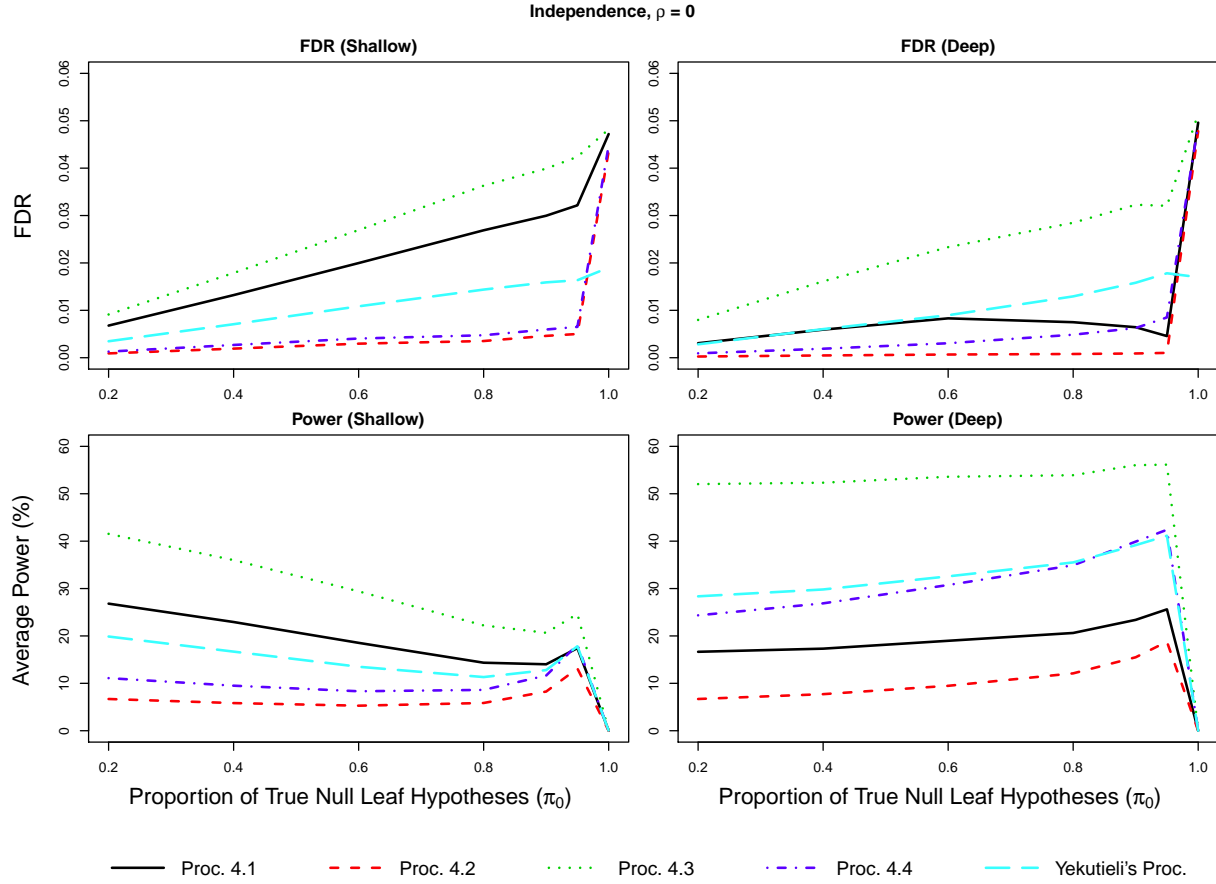
Figure 2: *FDR (top row) and average power (bottom row) of Procedures 1 (solid line), 2 (dashed), 3 (dotted), 4 (dot dash), and Yekutieli's procedure (long dash) under independence for the shallow hierarchy (left column) and the deep hierarchy (right column) where the proportion of true null leaf hypotheses varies from 0.2 to 1.*

obtained by averaging out the 5000 values of these two proportions, respectively.

Figure 2 displays the FDR and average power under independence as π_0 varies from 0.2 to 1. As seen from Figure 2, all the procedures control the FDR at level 0.05. In terms of power, Procedure 3 outperforms Yekutieli's procedure quite substantially and in some cases even doubles the power of Yekutieli's procedure. Procedure 1, which controls the FDR under positive dependence, outperforms Yekutieli's procedure under the shallow hierarchy but is outperformed by Yekutieli's procedure in the deep hierarchy. In the deep hierarchy, Procedure 4 and Yekutieli's procedure are comparable in terms of power. Not surprisingly, Procedure 2, which controls the FDR under arbitrary dependence, performs the worst.

Figures 3 and 4 display the FDR and average power under common correlation with $\rho = 0.25$ and $\rho = 0.75$, respectively, as π_0 varies from 0.2 to 1. The FDRs of all the pro-
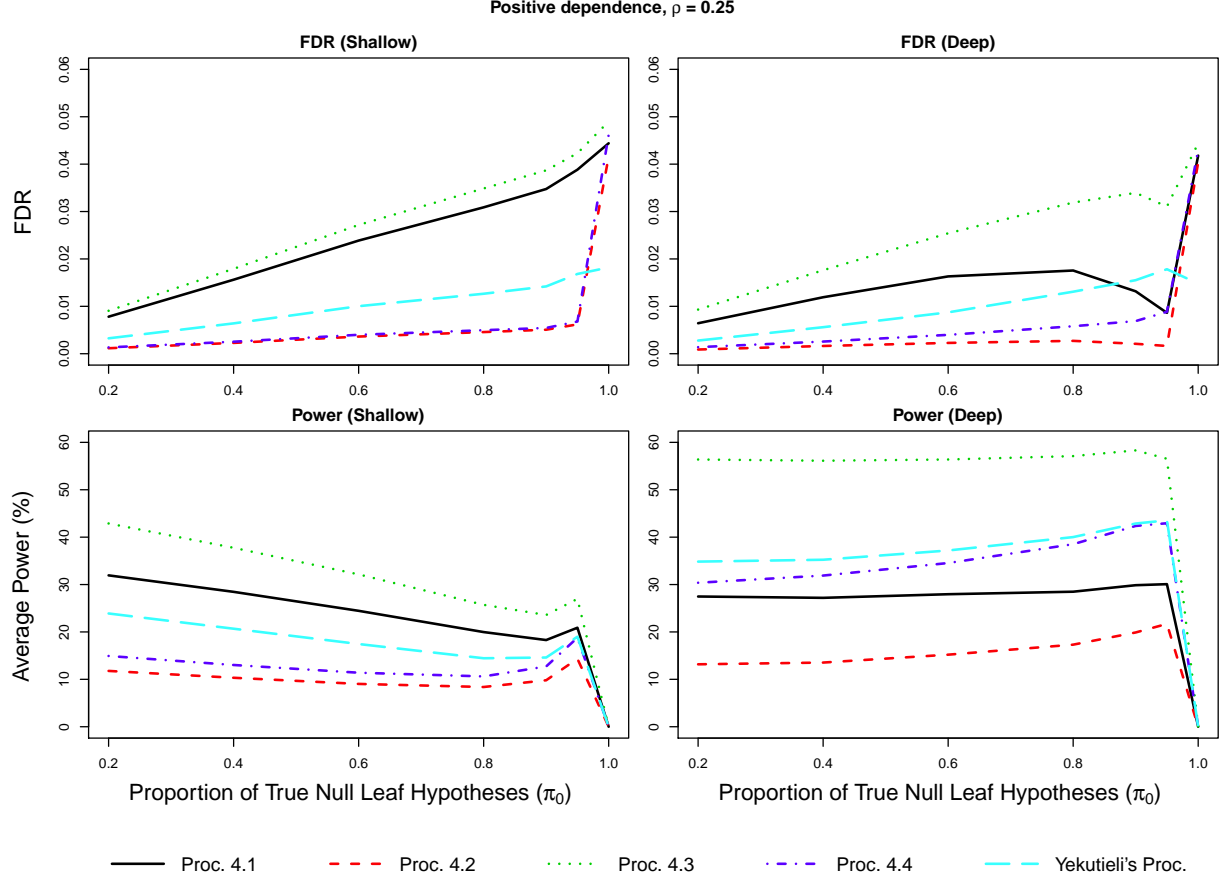
Figure 3: *FDR (top row) and average power (bottom row) of Procedures 1 (solid line), 2 (dashed), 3 (dotted), 4 (dot dash), and Yekutieli's procedure (long dash) under common correlation with $\rho = 0.25$ for the shallow hierarchy (left column) and the deep hierarchy (right column) where the proportion of true null leaf hypotheses varies from 0.2 to 1.*

cedures are controlled at level 0.05 under both weak and strong correlation. It should be noted that assumption 2 (block dependence) does not hold under this dependence configuration but Procedures 3 and 4 still control the FDR suggesting that both procedures are fairly robust to departures from this assumption. In terms of power, these figures show a similar pattern to Figure 2 where Procedure 3 is the most powerful and Procedure 2 is the least powerful. The remaining three procedures fall somewhere in the middle depending on the setting.

6. Real Data Analysis

We applied our proposed procedures as well as Yekutieli's procedure to a real data set. We used the data set of Caporaso et al. (2011), available in the phyoseq Bioconductor package
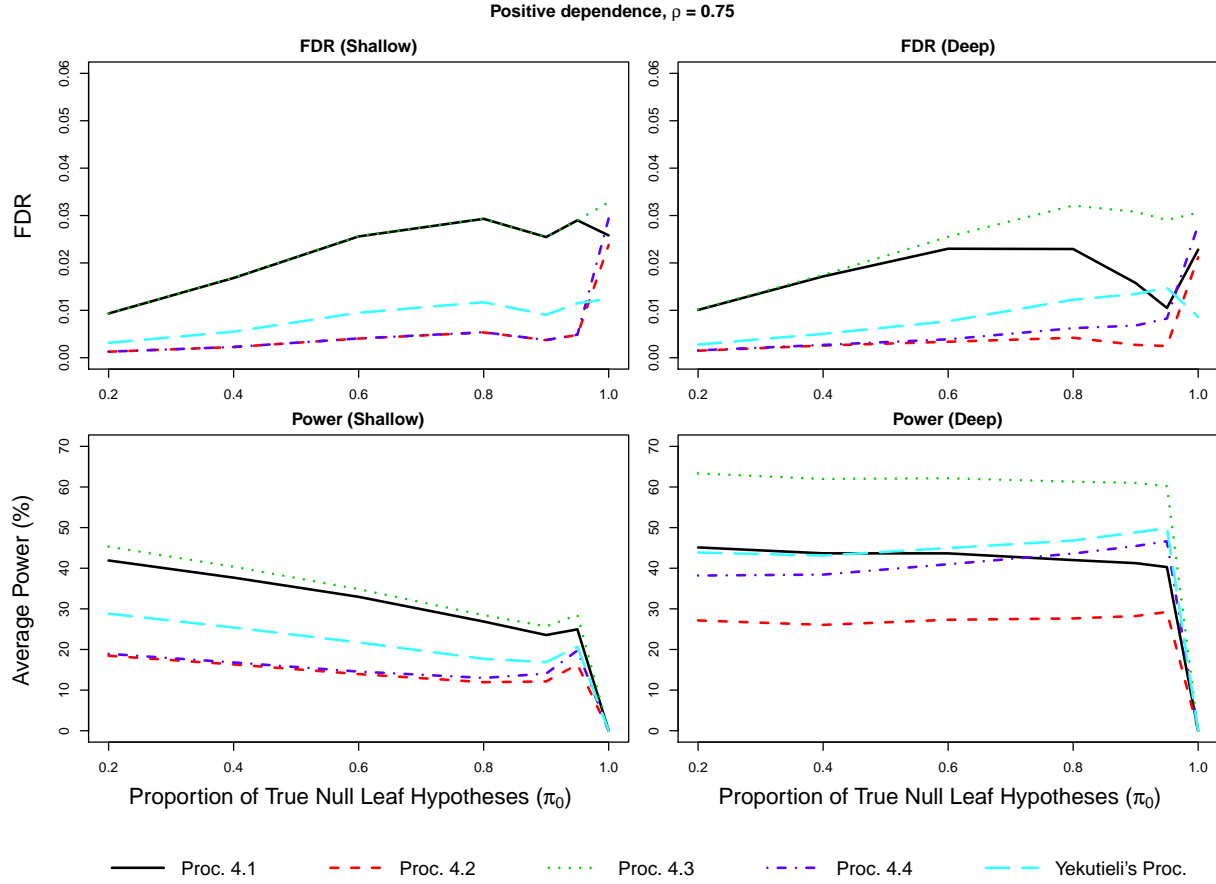
Figure 4: *FDR (top row) and average power (bottom row) of Procedures 1 (solid line), 2 (dashed), 3 (dotted), 4 (dot dash), and Yekutieli's procedure (long dash) under common correlation with $\rho = 0.75$ for the shallow hierarchy (left column) and the deep hierarchy (right column) where the proportion of true null leaf hypotheses varies from 0.2 to 1.*

at www.bioconductor.org, which provides the abundances of individual microbes in different ecological environments as well as their phylogenetic relationships. The data can be naturally organized into a hierarchy consisting of taxonomic units according to their phylogenetic relationships. The question of interest is whether there is an association between a taxonomic unit and ecological environment. Specifically, we tested the null hypothesis that the mean abundance for the taxonomic unit is the same across environments versus the alternative hypotheses that the mean abundance for the taxonomic unit is different across environments. The p-value for each hypothesis was determined by using an F-test where the abundance for a taxonomic unit in a given environment was determined based on the total abundance of each microbe within the taxonomic unit for the given environment (for more information see Sankaran and Holmes (2014)).

We restricted our analysis to the microbes in the Actinobacteria phylum which had 1631 individual microbes. The taxonomic hierarchy in the Actinobacteria phylum consisted of 3261 taxonomic units so that the total number of hypotheses is 3261 across 39 families. We tested the hypotheses at various significance levels and the number of rejections for each procedure are displayed in Table 5. All of the procedures are seen to make a substantial number of discoveries, even when $\alpha = 0.01$. In terms of the number of rejections, one can easily see that Procedure 3 is by far the best, significantly outperforming the other procedures. Procedure 4 outperforms Yekutieli's procedure when α is moderate to large but Yekutieli's procedure outperforms Procedure 4 when α is small. Procedure 2 is, not surprisingly, the worst since it is the only procedure that controls the FDR under arbitrary dependence.

Table 5: The number of rejections out of 3261 hypotheses by Procedures 1, 2, 3, 4, and Yekutieli's procedure at various significance levels for the microbe abundance data set of Caporaso et al. (2011) restricted to the Actinobacteria phylum.

α	Procedure 1	Procedure 2	Procedure 3	Procedure 4	Yekutieli's Procedure
0.01	75	68	144	107	123
0.025	88	75	574	148	165
0.05	118	92	1156	353	230
0.1	138	108	1497	813	253

7.   Conclusion

In this paper, we have developed several FDR controlling procedures for testing hierarchically ordered hypotheses. To our knowledge, we have, for the first time, presented hierarchical testing methods with proven FDR control under dependence. Furthermore, we have developed a method which controls the FDR under block positive dependence and in our simulation study, it was shown to be more powerful than Yekutieli's hierarchical testing procedure and other proposed procedures. A particularly interesting aspect of this work is that we have connected two contrasting testing methods in the proposed hierarchical testing methods: fixed sequence procedures, which assume the hypotheses have a fixed pre-defined testing order, and stepwise procedures, which do not assume the hypotheses having any pre-defined testing order.

We believe in this paper we have made a significant step in terms of multiple testing

with structured hypotheses. The techniques developed in this paper can be used to develop procedures to test hypotheses with more complex hierarchical structures where hypotheses are not restricted to only one parent. Such procedures would have applications towards testing interaction hypotheses, for example in gene expression data, where main effects are tested first and pairwise interactions are tested only if the two main effects making up the interaction are significant.

8. Appendix

Let us first state and prove the following lemmas which are used in the proofs of Theorems 1, 2, 3, and 4.

Lemma 1. *Under Assumption 1, if $\Gamma(P_1, \ldots, P_m)$ is a discrete coordinatewise non-increasing function of the p-values taking on values $\gamma_1 < \cdots < \gamma_n$ and $t(\cdot)$ is a non-decreasing function on $\{\gamma_1, \ldots, \gamma_n\}$, then for each true null H_j,*

$$\sum_{i=1}^{n} \Pr\left(\Gamma = \gamma_i \mid P_j \leq t(\gamma_i)\right) \leq \Pr\left(\Gamma \geq \gamma_1 \mid P_j \leq t(\gamma_1)\right).$$

Proof of Lemma 1.

$$\sum_{i=1}^{n} \Pr\left(\Gamma = \gamma_i \mid P_j \leq t(\gamma_i)\right)$$

$$= \sum_{i=1}^{n} \Pr\left(\Gamma \geq \gamma_i \mid P_j \leq t(\gamma_i)\right) - \sum_{i=1}^{n-1} \Pr\left(\Gamma \geq \gamma_{i+1} \mid P_j \leq t(\gamma_i)\right)$$

$$= \Pr\left(\Gamma \geq \gamma_1 \mid P_j \leq t(\gamma_1)\right) - \sum_{i=2}^{n} \left[\Pr\left(\Gamma \geq \gamma_i \mid P_j \leq t(\gamma_{i-1})\right) - \Pr\left(\Gamma \geq \gamma_i \mid P_j \leq t(\gamma_i)\right)\right]$$

$$\leq \Pr\left(\Gamma \geq \gamma_i \mid P_j \leq t(\gamma_1)\right).$$

The inequality follows by Assumption 1. ∎

Lemma 2. *Under arbitrary dependence of the p-values, if $\Gamma(P_1, \ldots, P_m)$ is a discrete function of the p-values taking on values $\gamma_1 < \cdots < \gamma_n$ and $t(\cdot)$ is a positive non-decreasing function on $\{\gamma_1, \ldots, \gamma_n\}$ with the convention that $t(\gamma_0) = 0$, then for each true null H_j,*

$$\sum_{i=1}^{n} \frac{1}{t(\gamma_i)} \Pr\left(\Gamma = \gamma_i, P_j \leq t(\gamma_i)\right) \leq \sum_{i=1}^{n} \frac{t(\gamma_i) - t(\gamma_{i-1})}{t(\gamma_i)}.$$

Proof of Lemma 2. Using the convention that $0/0 = 0$, we have

$$\sum_{i=1}^{n} \frac{1}{t(\gamma_i)} \Pr(\Gamma = \gamma_i, P_j \leq t(\gamma_i))$$

$$= \sum_{i=1}^{n} \left[\frac{1}{t(\gamma_i)} \Pr(\Gamma \geq \gamma_i, P_j \leq t(\gamma_i)) - \frac{1}{t(\gamma_{i-1})} \Pr(\Gamma \geq \gamma_i, P_j \leq t(\gamma_{i-1})) \right]$$

$$\leq \sum_{i=1}^{n} \frac{1}{t(\gamma_i)} \Pr(\Gamma \geq \gamma_i, t(\gamma_{i-1}) < P_j \leq t(\gamma_i))$$

$$\leq \sum_{i=1}^{n} \frac{1}{t(\gamma_i)} \Pr(t(\gamma_{i-1}) < P_j \leq t(\gamma_i))$$

$$= \sum_{i=1}^{n-1} \left(\frac{1}{t(\gamma_i)} - \frac{1}{t(\gamma_{i+1})} \right) \Pr(P_j \leq t(\gamma_i)) + \frac{1}{t(\gamma_n)} \Pr(P_j \leq t(\gamma_n))$$

$$\leq \sum_{i=1}^{n} \frac{t(\gamma_i) - t(\gamma_{i-1})}{t(\gamma_i)}. \blacksquare$$

8.1. Proof of Proposition 1. Assume $k > \psi(k)$. Then, step 2(a) of Algorithm 3.1 is repeated until for some $\ell \geq 2$, $r_\ell \leq \psi(r_\ell)$. For $t = 1, \ldots, \ell - 1$, we have $r_t > \psi(r_t)$ implying $r_t > r_{t+1}$. Thus, $r_\ell < r_1 = k$. For any integer r from 0 to $k - 1$ such that $r \leq \psi(r)$, we will show that $r_\ell \geq r$. To prove it, we show using induction that $r_t \geq r, t = 1, \ldots, \ell$. Since $r_1 = k > r$, by induction assume $r_{t-1} \geq r$. Then, $r_t = \psi(r_{t-1}) \geq \psi(r) \geq r$. Since $r_\ell \leq \psi(r_\ell)$, $r_\ell < k$, and $r_\ell \geq r$, we have $R = r_\ell = \max\{0 \leq r \leq k - 1 : r \leq \psi(r)\}$.

Conversely, assume $k \leq \psi(k)$. Then, step 2(b) is repeated until for some $\ell \geq 2$, $r_\ell > \psi(r_\ell)$. For $t = 1, \ldots, \ell - 1$, we have $r_t \leq \psi(r_t)$ implying $r_t < \psi(r_t) + 1 = r_{t+1}$. Thus, $r_\ell > r_1 = k$. For any integer r from $k + 1$ to $m + 1$ such that $r > \psi(r)$, we will show that $r_\ell \leq r$. To prove it, we show using induction that $r_t \leq r, t = 1, \ldots, \ell$. Since $r_1 = k < r$, by induction assume $r_{t-1} \leq r$. Then, $r_t = \psi(r_{t-1}) + 1 \leq \psi(r) + 1 \leq r$. Since $r_\ell > \psi(r_\ell)$, $r_\ell > k$, and $r_\ell \leq r$, we have $R = r_\ell - 1 = \min\{k + 1 \leq r \leq m + 1 : r > \psi(r)\} - 1$. $\blacksquare$

8.2. Proof of Theorem 1. In this proof and the remaining proofs, we will use the convention that $0/0 = 0$. For convenience of notation, define $\mathcal{G}_d = \bigcup_{j=1}^{d} \mathcal{F}_j$ and $R(\mathcal{G}_d)$ is the number of rejections in the first d families, $\mathcal{F}_j, j = 1, \ldots, d$. Let $|\mathcal{G}_d|$ be the cardinality of $\mathcal{G}_d$.

We will show that

$$E\left(\frac{V(\mathcal{M}_i)}{R} \right) \leq \frac{\ell_i \alpha}{\ell}, i = 1, \ldots, m. \tag{12}$$

Proof of (12). The event $\{H_i$ is rejected$\}$ implies all ancestors of H_i are rejected so there must be at least d_i rejections in the first d_i families. Therefore, the event $\{H_i$ is rejected$\}$ implies the following two inequalities:

$$d_i \leq R(\mathcal{G}_{d_i}) \leq |\mathcal{G}_{d_i}|, \tag{13}$$

$$R(\mathcal{G}_{d_i}) - 1 \leq R - R(\mathcal{M}_i). \tag{14}$$

The second inequality follows from the fact that $\mathcal{G}_{d_i}/\{H_i\} \subseteq \mathcal{M}/\mathcal{M}_i$ so that $R(\mathcal{G}_{d_i}/\{H_i\}) \leq R(\mathcal{M}/\mathcal{M}_i)$.

If H_i is true,

$$
\begin{aligned}
E\left(\frac{V(\mathcal{M}_i)}{R}\right) &\leq E\left(\frac{V(\mathcal{M}_i)}{V(\mathcal{M}_i) + R - R(\mathcal{M}_i)}\right) \\
&\leq E\left(\frac{m_i}{m_i + R - R(\mathcal{M}_i)} I\{H_i \text{ is rejected}\}\right) \\
&\leq E\left(\frac{m_i}{m_i + R(\mathcal{G}_{d_i}) - 1} I\{H_i \text{ is rejected}\}\right) \\
&= \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} E\left(\frac{m_i}{m_i + r - 1} I\{R(\mathcal{G}_{d_i}) = r, H_i \text{ is rejected}\}\right) \\
&\leq \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \frac{m_i}{m_i + r - 1} \Pr\left(R(\mathcal{G}_{d_i}) = r, P_i \leq \alpha_i(r)\right) \\
&\leq \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \frac{m_i \alpha_i(r)}{m_i + r - 1} \Pr\left(R(\mathcal{G}_{d_i}) = r \mid P_i \leq \alpha_i(r)\right) \\
&= \frac{\ell_i \alpha}{\ell} \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \Pr\left(R(\mathcal{G}_{d_i}) = r \mid P_i \leq \alpha_i(r)\right).
\end{aligned}
\tag{15}
$$

The second inequality follows from the fact that $V(\mathcal{M}_i) \leq m_i$ and $V(\mathcal{M}_i)/(V(\mathcal{M}_i) + R - R(\mathcal{M}_i))$ is an increasing function of $V(\mathcal{M}_i)$. The third inequality follows from (14) and the first equality follows from (13). The fourth inequality follows by the fact that the event $\{H_i$ is rejected$\} = \{H_{T(i)}$ is rejected, $P_i \leq \alpha_i(R(\mathcal{G}_{d_i}))\}$.

Since the number of rejections by the generalized stepup procedure is a coordinatewise non-increasing function of the p-values, it follows that $R(\mathcal{G}_{d_i})$ is also a coordinatewise non-

increasing function of the p-values. Therefore, by Lemma 1,

$$\sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \Pr\left(R(\mathcal{G}_{d_i}) = r \mid P_i \le \alpha_i(r)\right) \le \Pr\left(R(\mathcal{G}_{d_i}) \ge d_i \mid P_i \le \alpha_i(d_i)\right) \le 1. \quad (16)$$

From (16), we have that (15) is less than $\ell_i \alpha / \ell$. Thus, (12) holds when H_i is true.

We will use induction to show (12) also holds when H_i is false. When H_i is a false null leaf hypothesis, then (12) is true trivially. Otherwise, assume (12) is true for every false child hypothesis of H_i. Thus, (12) is true for all children of H_i. We note that when H_i is false, $V(\mathcal{M}_i) = \sum_{j:T(j)=i} V(\mathcal{M}_j)$ and

$$E\left(\frac{V(\mathcal{M}_i)}{R}\right) = \sum_{j:T(j)=i} E\left(\frac{V(\mathcal{M}_j)}{R}\right) \le \sum_{j:T(j)=i} \frac{\ell_j \alpha}{\ell} = \frac{\ell_i \alpha}{\ell}.$$

Thus, (12) holds for all true and false null hypotheses.

Proof of Theorem 1. By (12), we have

$$\mathrm{FDR} = \sum_{i:T(i)=0} E\left(\frac{V(\mathcal{M}_i)}{R}\right) \le \sum_{i:T(i)=0} \frac{\ell_i \alpha}{\ell} = \alpha. \quad \blacksquare$$

8.3. Proof of Theorem 2. We will show that (12) holds under arbitrary dependence for the procedure introduced in Theorem 2.

Proof of (12). When H_i is true, by the fourth inequality of (15), we have

$$E\left(\frac{V(\mathcal{M}_i)}{R}\right) \le \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \frac{m_i}{m_i + r - 1} \Pr\left(R(\mathcal{G}_{d_i}) = r, P_i \le \alpha_i(r)\right)$$

$$= \frac{\ell_i \alpha}{\ell} \frac{1}{c_i} \sum_{r=d_i}^{|\mathcal{G}_{d_i}|} \frac{1}{\alpha_i(r)} \Pr\left(R(\mathcal{G}_{d_i}) = r, P_i \le \alpha_i(r)\right)$$

$$\le \frac{\ell_i \alpha}{\ell} \frac{1}{c_i} \left(1 + \sum_{r=d_i+1}^{|\mathcal{G}_{d_i}|} \frac{\alpha_i(r) - \alpha_i(r-1)}{\alpha_i(r)}\right)$$

$$= \frac{\ell_i \alpha}{\ell}.$$

The second inequality follows by Lemma 2. Thus, (12) holds when H_i is true. When H_i is false, (12) also holds by the same argument used in the proof of Theorem 1. Hence, (12) holds

27

for all hypotheses.

Proof of Theorem 2. Since (12) holds for each $i = 1, \ldots, m$, FDR control follows by the same argument used in the proof of Theorem 1. ∎

8.4. Proof of Theorem 3. Recursively define the random variables $A_1, \ldots, A_m$ as follows:

$$
A_i = \begin{cases}
1 & T(i) = 0, \\
A_{T(i)} & T(i) \neq 0 \text{ and } H_{T(i)} \text{ is false}, \\
A_{T(i)}(1 - (1/R(\mathcal{G}_{d_{T(i)}})))I\{H_{T(i)} \text{ is rejected}\} & T(i) \neq 0 \text{ and } H_{T(i)} \text{ is true}.
\end{cases}
$$

Notice that A_i is a function of the p-values corresponding to the hypotheses in families $\mathcal{F}_1, \ldots, \mathcal{F}_{d_i-1}$ so that P_i and A_i are independent due to Assumption 2.

When H_i is a true null hypothesis, we have the following useful inequality

$$
E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))} \right) \leq E\left(A_i \right). \tag{17}
$$

Proof of (17). With the convention $R(\mathcal{G}_0) = 0$, we have

$$
\begin{aligned}
E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))} \right) &= E\left(A_i \frac{I\{H_{T(i)} \text{ is rejected}, P_i \leq \alpha_i(R(\mathcal{G}_{d_i}))\}}{\alpha_i(R(\mathcal{G}_{d_i}))} \right) \\
&= E\left(A_i \frac{I\{H_{T(i)} \text{ is rejected}, P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1} + R(\mathcal{F}_{d_i})))\}}{\alpha_i(R(\mathcal{G}_{d_i-1}) + R(\mathcal{F}_{d_i}))} \right) \\
&\leq E\left(A_i \sum_{r=1}^{|\mathcal{F}_{d_i}|} E\left(\frac{I\{P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1}) + r), R(\mathcal{F}_{d_i}) = r\}}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)} \,\middle|\, R(\mathcal{G}_{d_i-1}), A_i \right) \right) \\
&\leq E\left(A_i \sum_{r=1}^{|\mathcal{F}_{d_i}|} \Pr\left(R(\mathcal{F}_{d_i}) = r \mid P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1}) + r), R(\mathcal{G}_{d_i-1}), A_i \right) \right) \\
&\leq E\left(A_i \Pr\left(R(\mathcal{F}_{d_i}) \geq 1 \mid P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1}) + 1), R(\mathcal{G}_{d_i-1}), A_i \right) \right) \\
&\leq E\left(A_i \right).
\end{aligned}
$$

The first equality follows by the fact that the event $\{H_i \text{ is rejected}\}$ is equivalent to the event $\{H_{T(i)} \text{ is rejected}, P_i \leq \alpha_i(R(\mathcal{G}_{d_i}))\}$. The second inequality follows by the fact that P_i is independent of $R(\mathcal{G}_{d_i-1})$ and A_i due to Assumption 2 so that (4) still holds. Finally, the third

28

inequality is due to Lemma 1.

Now, we will show that

$$E\left(A_i\frac{V(\mathcal{M}_i)}{R}\right) \leq \frac{\ell_i\alpha}{\ell}E\left(A_i\right). \tag{18}$$

Proof of (18). If H_i is a false null leaf hypothesis, then the left hand side of (18) is 0. If H_i is a true null leaf hypothesis, then

$$E\left(A_i\frac{V(\mathcal{M}_i)}{R}\right) \leq E\left(A_i\frac{I\{H_i \text{ is rejected}\}}{R(\mathcal{G}_{d_i})}\right)$$
$$= \frac{\ell_i\alpha}{\ell}E\left(A_i\frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right) \leq \frac{\ell_i\alpha}{\ell}E\left(A_i\right).$$

The first inequality follows by the fact that $R(\mathcal{G}_{d_i}) \leq R$ and $V(\mathcal{M}_i) = I\{H_i \text{ is rejected}\}$ when H_i is a true null leaf hypothesis. The equality follows by $\alpha_i(r) = r\alpha/\ell$ and $\ell_i = 1$. The second inequality follows by (17). Thus, (18) holds when H_i is a leaf hypothesis.

Now, we will show that (18) holds when H_i is a non-leaf hypothesis. By induction assume (18) holds for all children of H_i. If H_i is false, then we note that $V(\mathcal{M}_i) = \sum_{j:T(j)=i} V(\mathcal{M}_j)$ and $A_i = A_j$ for each j such that $T(j) = i$. Thus,

$$E\left(A_i\frac{V(\mathcal{M}_i)}{R}\right) = E\left(\sum_{j:T(j)=i} A_j\frac{V(\mathcal{M}_j)}{R}\right)$$
$$\leq \sum_{j:T(j)=i} \frac{\ell_j\alpha}{\ell}E\left(A_j\right) = \frac{\ell_i\alpha}{\ell}E\left(A_i\right).$$

The inequality follows by induction.

Now, assume H_i is true. We will use the following inequality,

$$\frac{1}{R} = \frac{1}{R(\mathcal{G}_{d_i})} - \frac{R - R(\mathcal{G}_{d_i})}{RR(\mathcal{G}_{d_i})}$$
$$\leq \frac{1}{R(\mathcal{G}_{d_i})} - \sum_{j:T(j)=i} \frac{R(\mathcal{M}_j)}{RR(\mathcal{G}_{d_i})}$$
$$\leq \frac{1}{R(\mathcal{G}_{d_i})} - \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{RR(\mathcal{G}_{d_i})}. \tag{19}$$

The equality follows by simple algebra and the first inequality follows by the fact that $\mathcal{M}_j \subseteq$

$\mathcal{M}/\mathcal{G}_{d_i}$ for each j with $T(j) = i$ so that $\sum_{j:T(j)=i} R(\mathcal{M}_j) \leq R - R(\mathcal{G}_{d_i})$. The second inequality follows by the fact that $R(\mathcal{M}_j) \geq V(\mathcal{M}_j)$ for each j. It should also be noted that $V(\mathcal{M}_i) = (1 + \sum_{j:T(j)=i} V(\mathcal{M}_j))I\{H_i \text{ is rejected}\}$. Thus,

$$
\begin{aligned}
& E\left(A_i \frac{V(\mathcal{M}_i)}{R}\right) \\
= \ & E\left(A_i \left(\frac{1}{R} + \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{R}\right) I\{H_i \text{ is rejected}\}\right) \\
\leq \ & E\left(A_i \left(\frac{1}{R(\mathcal{G}_{d_i})} - \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{RR(\mathcal{G}_{d_i})} + \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{R}\right) I\{H_i \text{ is rejected}\}\right) \\
= \ & E\left(A_i \left(\frac{1}{R(\mathcal{G}_{d_i})} + \left(1 - \frac{1}{R(\mathcal{G}_{d_i})}\right) \sum_{j:T(j)=i} \frac{V(\mathcal{M}_j)}{R}\right) I\{H_i \text{ is rejected}\}\right) \\
= \ & E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{R(\mathcal{G}_{d_i})} + \sum_{j:T(j)=i} A_j \frac{V(\mathcal{M}_j)}{R}\right) \\
\leq \ & E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{R(\mathcal{G}_{d_i})} + \sum_{j:T(j)=i} \frac{\ell_j \alpha}{\ell} A_j\right) \\
= \ & E\left(A_i \left(\frac{1}{R(\mathcal{G}_{d_i})} + \left(1 - \frac{1}{R(\mathcal{G}_{d_i})}\right) \frac{\ell_i \alpha}{\ell}\right) I\{H_i \text{ is rejected}\}\right) \\
= \ & \frac{\ell_i \alpha}{\ell} E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right) \\
\leq \ & \frac{\ell_i \alpha}{\ell} E\left(A_i\right).
\end{aligned}
$$

The first inequality follows by (19). The third and forth equality follow by the fact that $A_j = A_i(1 - (1/R(\mathcal{G}_{d_i}))I\{H_i \text{ is rejected}\}$ for j such that $T(j) = i$, since H_i is true. The second inequality follows by induction. The last equality follows by $\alpha_i(r) = \ell_i r \alpha / (\ell + \ell_i(r-1)\alpha)$ and the last inequality follows by (17).

Proof of Theorem 3. Finally, by (18),

$$
\begin{aligned}
\text{FDR} = E\left(\frac{V}{R}\right) &= \sum_{i:T(i)=0} E\left(\frac{V(\mathcal{M}_i)}{R}\right) \\
&= \sum_{i:T(i)=0} E\left(A_i \frac{V(\mathcal{M}_i)}{R}\right) \leq \sum_{i:T(i)=0} \frac{\ell_i \alpha}{\ell} = \alpha. \ \blacksquare
\end{aligned}
$$

8.5. Proof of Theorem 4.

In the proof, we will use the same notations as in the proof of Theorem 3. To prove Theorem 4, we show that the following inequality holds when H_i is true

$$E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right) \leq c_i E(A_i). \tag{20}$$

Proof of (20). It can be easily shown through simple algebra that for any leaf or non-leaf hypothesis H_i, the constant c_i can be expressed as

$$c_i = 1 + \sum_{r=2}^{|\mathcal{F}_{d_i}|} \frac{\alpha_i(r + d_i - 1) - \alpha_i(r + d_i - 2)}{\alpha_i(r + d_i - 1)}. \tag{21}$$

Assume H_i is true. Then,

$$E\left(A_i \frac{I\{H_i \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right)$$

$$= E\left(A_i \frac{I\{P_i \leq \alpha_i(R(\mathcal{G}_{d_i})), H_{T(i)} \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i}))}\right)$$

$$= E\left(A_i \frac{I\{P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1}) + R(\mathcal{F}_{d_i})), H_{T(i)} \text{ is rejected}\}}{\alpha_i(R(\mathcal{G}_{d_i-1}) + R(\mathcal{F}_{d_i}))}\right)$$

$$= E\left(A_i I\{H_{T(i)} \text{ is rejected}\} \times \right.$$

$$\left. E\left(\sum_{r=1}^{|\mathcal{F}_{d_i}|} \frac{I\{P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1}) + r), R(\mathcal{F}_{d_i}) = r\}}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)} \,\middle|\, \mathbf{P}_{d_i-1}\right)\right)$$

$$= E\left(A_i I\{H_{T(i)} \text{ is rejected}\} \times \right.$$

$$\left. \sum_{r=1}^{|\mathcal{F}_{d_i}|} \frac{\Pr\left(P_i \leq \alpha_i(R(\mathcal{G}_{d_i-1}) + r), R(\mathcal{F}_{d_i}) = r \mid \mathbf{P}_{d_i-1}\right)}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)}\right)$$

$$\leq E\left(A_i I\{H_{T(i)} \text{ is rejected}\} \left(1 + \sum_{r=2}^{|\mathcal{F}_{d_i}|} \frac{\alpha_i(R(\mathcal{G}_{d_i-1}) + r) - \alpha_i(R(\mathcal{G}_{d_i-1}) + r - 1)}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)}\right)\right)$$

$$\leq E\left(A_i I\{R(\mathcal{G}_{d_i-1}) \geq d_i - 1\} \left(1 + \sum_{r=2}^{|\mathcal{F}_{d_i}|} \frac{\alpha_i(R(\mathcal{G}_{d_i-1}) + r) - \alpha_i(R(\mathcal{G}_{d_i-1}) + r - 1)}{\alpha_i(R(\mathcal{G}_{d_i-1}) + r)}\right)\right)$$

$$\leq E\left(A_i \left(1 + \sum_{r=2}^{|\mathcal{F}_{d_i}|} \frac{\alpha_i(r + d_i - 1) - \alpha_i(r + d_i - 2)}{\alpha_i(r + d_i - 1)}\right)\right)$$

$$= c_i E\left(A_i\right).$$

Here, $\mathbf{P}_{d_i-1}$ denotes the p-value vector consisting of the p-values corresponding to the hypotheses in the first $d_i - 1$ families, $\mathcal{F}_1, \ldots, \mathcal{F}_{d_i-1}$. The first equality follows by the fact that the event $\{H_i \text{ is rejected}\}$ is equivalent to the event $\{H_{T(i)} \text{ is rejected}, P_i \leq \alpha_i(R(\mathcal{G}_{d_i}))\}$. The first inequality follows by Lemma 2 and the fact that P_i is independent of $\mathbf{P}_{d_i-1}$ due to Assumption 2 and $R(\mathcal{G}_{d_i-1})$ is determined by $\mathbf{P}_{d_i-1}$. The second inequality follows by the fact that the event $\{H_{T(i)} \text{ is rejected}\}$ implies all ancestors of $H_{T(i)}$ are rejected, so there must be at least $d_i - 1$ rejections in the first $d_i - 1$ families, i.e., $R(\mathcal{G}_{d_i-1}) \geq d_i - 1$. The third inequality follows by the fact that $[\alpha_i(R(\mathcal{G}_{d_i-1}) + r) - \alpha_i(R(\mathcal{G}_{d_i-1}) + r - 1)]/\alpha_i(R(\mathcal{G}_{d_i-1}) + r)$ is a decreasing function of $R(\mathcal{G}_{d_i-1})$ for each given r. The last equality follows from (21).

Proof of Theorem 4. By using the same argument for the proof of (18), we have that

$$E\left(A_i V(\mathcal{M}_i)/R\right) \leq \ell_i \alpha E\left(A_i\right)/\ell.$$

Thus, the FDR control of this procedure follows by the same argument used in the proof of Theorem 3. ∎

References

Benjamini, Y., and Heller, R. (2007), "False discovery rates for spatial signals," *J. Amer. Satist. Assoc.*, 102, 1272–1281.

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

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

Brechenmacher, T., Xu, J., Dmitrienko, A., and Tamhane, A. (2011), "A mixture gatekeeping procedure based on the Hommel test for clinical trial applications," *Journal of Biopharmaceutical Statistics*, 21, 748–767.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., and Knight, R. (2011), "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample," *PNAS*, 108 Suppl 1, 4516–4522.

Dmitrienko, A., Offen, W., Wang, O., and Xiao, D. (2006), "Gatekeeping procedures in dose-response clinical trials based on the Dunnet test," *Pharmaceutical Statistics*, 5, 19–28.

Dmitrienko, A., Wiens, B., Tamhane, A., and Wang, X. (2007), "Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives," *Statistics in Medicine*, 26, 2465–2478.

Goeman, J., and Finos, L. (2012), "The inheritance procedure: multiple testing of tree-structured hypotheses," *Statistical Applications in Genetics and Molecular Biology*, 11(1), 1–18.

Goeman, J., and Mansmann, U. (2008), "Multiple testing on the directed acyclic graph of gene ontology," *Bioinformatics*, 24, 537–544.

Guo, W., and Sarkar, S. (2016), "Adaptive controls of FWER and FDR under block dependence," *arXiv preprint arXiv:1611.03155*, .

Guo, W., Sarkar, S., and Peddada, S. (2010), "Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories," *Bioinformatics*, 66, 485–492.

Heller, R., Manduchi, E., Grant, G., and Ewens, W. (2009), "A flexible two-stage procedure for identifying gene sets that are differentially expressed," *Bioinformatics*, 25, 929–942.

Huque, M., and Alosh, M. (2008), "A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials," *Journal of Statistical Planning and Inference*, 138, 321–335.

Lynch, G., Guo, W., Sarkar, S., and Finner, H. (2016), "The control of the false discovery rate in fixed sequence multiple testing," *arXiv preprint arXiv:1611.03146*, .

Mehrotra, D., and Heyse, J. (2004), "Use of the false discovery rate for evaluating clinical safety data," *Statistical Methods in Medical Research*, 13, 227–238.

Meinshausen, N. (2008), "Hierarchical testing of variable importance," *Biometrika*, 95, 265–278.

Sankaran, K., and Holmes, S. (2014), "structSSI: simultaneous and selective inference for hierarchically structured data," *Journal of Statistical Software*, 59, 1–21.

Sarkar, S. (1998), "Some probability inequalities for ordered MTP 2 random variables: a proof of the Simes conjecture," *Ann. Statist.*, 26, 494–504.

Sarkar, S. (2002), "Some results on false discovery rate in stepwise multiple testing procedures," *Ann. Statist.*, 30, 239–257.

Tamhane, A., Liu, W., and Dunnet, C. (1998), "A generalized step-up-down multiple test procedure," *The Candian Journal Of Statistics*, 26, 353–363.

Yekutieli, D. (2008), "Hierarchical false discovery rate-control methodology," *J. Amer. Satist. Assoc.*, 103, 309–316.

Yekutieli, D., Reiner-Benaim, A., and Benjamini, Y. (2006), "Approaches to multiplicity issues in complex research in microarray analysis," *Statistica Neerlandica*, 60, 414–437.