Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Yunxiao Li	Date

DEVELOPMENT OF STATISTICAL METHODS FOR MULTIPLE-HYPOTHESES TESTING

By

Yunxiao Li

Doctor of Philosophy

Biostatistics

Yijuan Hu, Ph.D.
Advisor

Glen A. Satten, Ph.D.
Co-advisor

Zhaohui Steve Qin, Ph.D.
Committee Member

Tianwei Yu, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

DEVELOPMENT OF STATISTICAL METHODS FOR MULTIPLE-HYPOTHESES TESTING

By

Yunxiao Li

M.S., Emory University, 2018

B.S., Peking University, 2014

Advisors: Yijuan Hu, Ph.D. and Glen A. Satten, Ph.D.

An Abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics and Bioinformatics

2019

Abstract

Development of Statistical Methods for Multiple-Hypotheses Testing

By Yunxiao Li

In this dissertation, I develop three novel statistical methods for solving multiple-hypotheses testing problems.

In the first topic, we propose a bottom-up approach to testing hypotheses that have a tree-structured dependency structure. Our motivating example comes from testing the associations between a trait of interest and groups of microbes that have been organized into operational taxonomic units (OTUs). Given p-values from association tests for each individual OTU, we would like to know if we can declare that a certain species, genus, or higher taxonomic group can be considered to be associated with the trait. For this problem, a bottom-up testing algorithm that starts at the lowest level of the tree (OTUs) and proceeds upward through successively higher taxonomic groups (species, genus, family etc.) is required. We develop such a bottom-up testing algorithm that controls the error rate of decisions made at higher levels in the tree, conditioning on findings at lower levels in the tree. We further show that our algorithm controls the false discovery rate based on the global null hypothesis that no taxa are associated with the trait. By simulation, we also show that our approach is better at finding driver taxa, the highest level taxa below which there are dense association signals. We illustrate our approach using data from a study of the microbiome among patients with ulcerative colitis and healthy controls.

In the second topic, we consider the resampling-based multiple testing problems. In multiple testing literature, the standard procedures for correcting multiplicity, such as the Benjamini and Hochberg (1995) procedure, usually require knowledge of the ideal p-values. In many biological applications such as microbiome studies, the ideal p-values cannot be computed analytically but only approximated by resampling methods. The resampling-based p-values coupled with a multiplicity-correction procedure generally produce different lists of rejections when the resampling algorithm is initiated by different random seeds, hence lacking of reproducibility. The existing method of Gandy and Hahn (2014, 2016) that aimed to control the Monte Carlo (MC) error (i.e., disagreement with the decisions based on ideal p-values) rate is extremely conservative and tends to make zero rejection. We focus on the type-I MC error which occurs when we reject a hypothesis that should be accepted according to the ideal decisions (in other words, this rejection is not reproducible). We develop a two-step algorithm based on resampling replicates to make decisions while controlling the type-I MC error rate. Through extensive simulation studies, we demonstrate substantial power improvement compared to the existing method.

In the third topic, we propose a class of algorithms for sequential resampling-based multiple testing. Resampling-based tests are known to be computationally intensive. Sequential algorithms provide efficient and accurate estimation to ideal p-values in resampling-based tests, by allowing early stopping of generating resampling samples as long as evidence suggests that a hypothesis should be classified to rejection or acceptance region. However, most existing sequential methods in this field (e.g., Sandve et al. (2011)) cannot guarantee reproducibility of test decisions. The only sequential methods that addressed this issue were developed by Gandy and Hahn (2014, 2016). We develop novel sequential testing algorithms by incorporating a step-wise decision process and improved sequential confidence intervals. Performances of our proposed methods are assessed through both synthetic and real data.

DEVELOPMENT OF STATISTICAL METHODS FOR MULTIPLE-HYPOTHESES TESTING

By

Yunxiao Li

M.S., Emory University, 2018

B.S., Peking University, 2014

Advisors: Yijuan Hu, Ph.D. and Glen A. Satten, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2019

Acknowledgments

I would like to express my most sincere appreciation to my advisors, Dr. Yijuan Hu and Dr. Glen A. Satten, for teaching me how to discover good research questions, for encouraging me to explore new research ideas, and for supporting me unconditionally during my Ph.D. studies at Emory. In addition to their knowledge and skills, the passion and enthusiasm that they have in their research have influenced me deeply. They are my life-time mentors.

I would like to thank my dissertation committee members, Dr. Zhaohui Steve Qin and Dr. Tianwei Yu, for their thoughtful comments and insightful suggestions. They helped me improve my dissertation work significantly.

I would like to thank Dr. Jeanie Park for the unique opportunity to collaborate with her team. This experience provides me with invaluable exposure to practical innovative research that moves science forward.

Last but not least, I would also like to thank my parents and my wife. This dissertation would not have been possible without their understanding, love and endless support.

Contents

List of Tables

List of Figures

Chapter 1

Background and Literature Review

Multiple-hypotheses testing refers to the situation where we test more than one hypothesis at the same time. In the classic problem of single-hypothesis testing, we usually find the distribution of test statistic, calculate a p-value that indicates significance level, and compare the p-value to a pre-specified threshold α (e.g., 5%). If p-value $\leq \alpha$, then we reject the null hypothesis H_0. This approach can control the type-I error rate $\Pr(\text{reject } H_0|H_0) \leq \alpha$. However, it does not control the type-I error rate in multiple testing problems. For example, we test m hypotheses $H_{1,0}, H_{2,0}, \cdots, H_{m,0}$ and p-values are $p_1, p_2, \cdots, p_m$. It could be a study for gene expression. Each hypothesis corresponds to a gene, and rejecting a hypothesis means that the gene is differentially expressed under two experimental conditions. We know that null p-values follow uniform distribution. If we assume independence among p_i, and the truth is that all hypotheses are under the null, the type-I error using this approach would be:

$$\Pr\left(\bigcup_{i=1}^{m}\{p_i \leq \alpha\}\right) = 1 - \Pr\left(\bigcap_{i=1}^{m}\{p_i > \alpha\}\right) = 1 - \prod_{i=1}^{m}\Pr\left(p_i > \alpha\right)$$
$$= 1 - (1-\alpha)^m \overset{m\to\infty}{\longrightarrow} 1.$$

In this case, we would almost surely reject at least one hypothesis, although any rejection here must be an error. By no means these 'significant' findings can be replicated by other laboratories or other researchers. New testing procedures must be developed to reduce those erroneous discoveries that could be possibly concluded by multiple testing.

The solution to the multiplicity issue is usually controlling a stronger error criterion in multiple testing. The first error rate of interest is called family-wise error rate (FWER) which equals the probability of falsely rejecting at least one null hypothesis among the entire family of hypotheses, or explicitly $\Pr(V \geq 1)$ using the notations from Table 1.1. The Bonferroni correction (Bonferroni, 1936) is the simplest way to control FWER. Basically the p-value in the example would be compared to α/m rather than α. However, the Bonferroni correction is known to be conservative, and usually only works for small m. When testing a large amount of hypotheses in genetics and genomics studies, false discovery rate (FDR) (Benjamini and Hochberg, 1995) would usually be the ideal error criterion. FDR controlling procedures uniformly yield higher statistical power than

FWER controlling procedures. FDR measures the average of the false discovery proportion (FDP). Using notations in Table 1.1, FDR = $\mathbb{E}(\text{FDP}) = \mathbb{E}(\frac{V}{R \vee 1})$, where V is called the number of false discoveries and R is called the number of total discoveries. The Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) is the most popular and influential FDR controlling approach in past decades.

Table 1.1: Summary of all possible outcomes when testing multiple hypotheses

	Acceptances	Rejections	Total
True null	U	V	m_0
Non-null	T	S	$m - m_0$
Total	$m - R$	R	m

Since early 1960s, a significant amount of statistical methods have been developed for multiple testing. However, there are still many open questions in this field. We discuss several important ones in this dissertation. First, in general it is difficult for a multiple testing method to achieve decent power while controlling the error rates under arbitrary correlation or dependency structure among test statistics. We consider a special scenario where multiple tree-structured hypotheses are tested. Our problem is particularly motivated by applications in human microbiome studies. In Section 1.1 we introduce more details of this problem. Our proposed solution is presented in Chapter 2. Second, many existing multiple testing procedures (e.g., the BH procedure) can still apply to p-values from resampling-based tests. However, uncertainty in resampling process usually causes variation in the decisions from multiple testing procedures. A decision can easily change even if we use a different random seed for resampling. We call this Monte Carlo (MC) error. More introduction of MC error can be found in Section 1.2 and 1.3. In Chapter 3, we propose powerful multiple testing procedures that provides theoretical guarantee on type-I MC error rate control. Sequential algorithms can be helpful to reduce the intensive computational cost of resampling-based multiple testing. The MC error also exists in sequential resampling-based multiple testing procedures. We review existing methods in Section 1.3, and propose novel sequential algorithms that address this problem in Chapter 4.

1.1 Testing Tree-Structured Hypotheses in Microbiome Data

The FDR criterion has largely replaced FWER as a benchmark for testing many hypotheses. Benjamini and Hochberg (1995) proposed a simple way to to control the FDR when testing independent hypotheses, which extends easily to hypotheses having a positive regression dependence structure (Benjamini and Yekutieli, 2001). More general conditions for BH to asymptotically control FDR were discussed by Farcomeni (2007) and Wu (2008). However, under the general dependency structure, Qiu et al. (2005), Efron (2007) and Schwartzman and Lin (2011) showed that correlation among test statistics may have a significant impact on FDR control. A log-factor adjustment to the Benjamini and Hochberg procedure (Benjamini and Yekutieli, 2001) allows FDR control under arbitrary correlation among tests. However, this procedure is very conservative in most settings.

Testing procedures that control FDR for specific patterns of dependence have also been investigated. Sun and Cai (2009) developed multiple testing methods by assuming that the dependency structure of test statistics can be characterized by a hidden Markov model. Multiple testing procedures under the factor model were studied by Leek and Storey (2008), Friguet et al. (2009), and Fan et al. (2012). Barber and Candès (2015) and Candès et al. (2018) developed knockoff-filter, a powerful method to control FDR when the dependency structure of test statistics is embedded in linear regression model.

In some problems, the dependency structure among test statistics can be characterized by a tree structure. Yekutieli (2008) considered hypotheses tests organized in a branching tree using an approach that starts by testing the hypothesis at the "top" of the tree; if this hypothesis is rejected, hypotheses at the next lowest level are tested. Testing continues from top to bottom until no further hypotheses can be rejected, at which point no further tests are conducted. This approach is appropriate for some problems, such as the motivating example in Yekutieli (2008) in which a genome-wide test for (genetic) linkage was conducted, followed by tests for linkage separately on each chromosome, then tests for linkage on the p or q arms of each chromosome, etc. Moving down the tree corresponds to increasing localization of the linkage signal, making the top-down strategy a natural choice. More top-down methods that test tree-structured hypotheses were developed by Meinshausen (2008) and Rosenbaum (2008) when the goal is to control

FWER; by Benjamini and Heller (2007), Benjamini and Bogomolov (2014), Peterson et al. (2016) and Heller et al. (2018) when the goal is to control FDR or its generalizations; and by Ramdas et al. (2017) for testing multiple hypotheses on directed acyclic graphs. For those top-down tests, the null hypothesis being tested at each node in the tree is usually the "global null hypothesis" which means that all tests below this node are under their null conditions respectively. Methodologies of testing global nulls mainly include the Fisher's method (Fisher, 1992), Stouffer's Z test (Stouffer et al., 1949), minimum p-value approach, Pearson's test (Pearson, 1938; Owen, 2009), and Sime's test (Simes, 1986). Loughin (2004) comprehensively reviewed these approaches and concluded that each method is only suitable and outperforms other methods in its own scenario. A global testing method that works under arbitrary correlations was developed by Pillai and Meng (2016) and Liu and Xie (2019), based on a nice property that the sum of Cauchy random variables approximately follows Cauchy distribution regardless of correlation.

Our motivating data application comes from human microbiome research. We will show that this example is not well-suited to the top-down approach. Microbiome refers to small living organisms inhabiting on human bodies such as bacteria, viruses, archaea, fungi and et cetera. Despite being so small, the population of human microbiome is huge: it was estimated that human microbiome contains at least 10-100 trillion cells, approximately 10-fold more than the number of human cells (Turnbaugh et al., 2007). The collective genome of human microbiome contains approximately 150 times more genes than the human genome (Qin et al., 2010). Recent studies have revealed important connections between microbiome and human health. Microbiome has been found to be associated with cardiovascular disease (Tang and Hazen, 2017), inflammatory bowel diseases (Morgan et al., 2012; Halfvarson et al., 2017), type-II diabetes (Qin et al., 2012) and pathophysiology of neurodevelopmental disorders (Hsiao et al., 2013). Microbiome can be profiled by 16S ribosomal RNA sequencing technology (Chakravorty et al., 2007). These 16S sequence reads can be grouped together with other reads from similar species or sub-species. Such a group structure is usually recognized as an operational taxonomic units (OTUs) or assigned to amplicon sequence variants (ASVs). For simplicity, we restrict our discussion to OTUs with the understanding that the argument can apply to either ASVs or OTUs. It is usually a natural scientific question to assess whether or

not a microbiome marker is associated with a trait of interest such as disease phenotype. Existing methods can either test the global composition among microbial communities (Anderson, 2001; Zhao et al., 2015; Hu and Satten, 2017), or identify individual OTUs that are associated with the trait (Robinson et al., 2010; Love et al., 2014; Paulson et al., 2013; Chen and Li, 2016; Fang et al., 2016; Hu and Satten, 2017).

After testing association on each OTU, we may wish to determine if any species of microbes are associated with the trait. Depending on our findings, we may wish to test larger groups corresponding to successively higher taxonomic ranks (e.g., genus, family, order, class, phylum, and kingdom). The natural ordering of hypotheses in this example starts at the bottom of the tree and proceeds upward. Further, it may be desirable to continue to test hypotheses at higher levels of the tree even if no findings have been made at a lower level, since an accumulation of weak signals from lower levels may coalesce into a detectable signal at a higher level. The scientific questions of interest thus motivate development of a bottom-up approach to testing tree-structured hypotheses. In literature only few methods (Tang et al., 2017, 2018) developed association testing procedures to detect subtrees under which 16S read counts are unbalanced between the case and control groups. However, these methods only take a naïve approach to address multiplicity issue, in which dependency among taxa is not accounted for.

When considering if we should declare a certain node (say, a genus) to be associated with the trait we are studying, we adopt the following approach: if a large proportion of species from that genus influence the trait, we should conclude the genus influences the trait. Conversely if only a few of the species from a genus are non-null, then a better description of the microbes that influence occurrence of the trait is a list of associated species. Finding taxa that can be said to influence a trait in this sense is the first goal of our approach. The second goal is to locate the highest taxa in the tree for which we can conclude many taxa below, but not any ancestors above, influence risk; we refer to such taxa as *driver* taxa. We also consider a related criterion, the *conjunction* null hypothesis, that would require that *all* species from the genus be associated with the trait before declaring the genus is associated. The concept of conjunction null was proposed by Nichols et al. (2005) and Friston et al. (2005). For a brain imaging application,

Nichols et al. (2005) proposed an approach that considers the minimum of multiple t-statistics. Equivalently the maximum p-value among all individual tests can be used as the p-value for testing conjunction null. However, this could be extremely conservative in most settings. A more liberal definition for null hypothesis is called partial conjunction null, which was proposed by Benjamini and Heller (2008).

1.2 Resampling-Based Multiple Testing and Monte Carlo Error

Modern scientific studies often involve testing multiple hypotheses, in which individual hypotheses are first tested and a procedure that corrects for multiplicity is then applied. For example, in many applications, BH (Benjamini and Hochberg, 1995) procedure can be used to control FDR. The Bonferroni (Bonferroni, 1936) or Holm's (Holm, 1979) procedure can control the FWER. In particular, the BH procedure has received most attention, due to its use of the FDR criterion that tends to promote more scientific discoveries.

The above multiplicity-correction procedures require knowledge of the *ideal* p-values of all tests. The ideal p-values can be analytically calculated when the distribution of test statistic (e.g., t-statistic) is simple. When the distribution of test statistic at each hypothesis is intractable, resampling-based test (e.g., permutation test) would be usually preferred. Resampling-based tests only require mild assumptions such as the exchangeability condition (De Finetti, 1992) for permutation tests. The resampling-based tests are known to be more robust than likelihood-based tests under small sample size. In principle, the ideal p-values of permutation tests can be obtained by running through all permutation replicates. We refer to the decisions, either *rejection* or *acceptance* for each test, based on the ideal p-values as ideal decisions. In many cases, the ideal p-values are not available but can only be approximated by resampling methods which take a small (compared to the enumerating all permutations or bootstrap replicates) random sample of the possible replicates. For ease of presentation, we call such a resampling replicate a permutation sample. A common practice is to collect a pre-specified number of permutations, compute permutation p-values as the proportion of permutation test statistics that are more "extreme" than the observed test statistic, and then apply one

of the multiplicity-correction procedures; we refer to this approach as the *naïve* decision rule. In general, it is a tricky question to choose an appropriate number of permutations. For bootstrap resampling, the choice of total replicates has been widely discussed by Hall (1986); Efron (1987); Davison and Hinkley (1997). Methods that enhance numerical performances of bootstrap procedures were developed by Gleason (1988); Davison et al. (1986); Hinkley (1988). However, the main focus is effectiveness (i.e., power and coverage) and computational efficiency of bootstrap methods. For permutation tests, it was suggested that at least $O(1/p^*)$ permutation samples are required for relatively accurate estimation of ideal p-value p^* (Kimmel and Shamir, 2006; Yu et al., 2011). This estimated range may vary dramatically. Followed by the $100/p^*$ rule mentioned in Yu et al. (2011), we would require as many as 10^9 replicates when the ideal p-value $= 10^{-7}$. It can be shown that the naïve decision rule with the standard BH procedure still controls FDR under permutation p-values as long as null permutation p-values are stochastically greater than U$[0, 1]$ (Benjamini and Hochberg, 1995). To satisfy this condition, usually at each test we pick the permutation p-value which can be written as

$$\frac{1 + \sum_{j=1}^{n} \mathbb{I}(T_j > t)}{n + 1},$$

where n is the number of permutation samples, t is the test statistic from the observed data, and T_j represents the permuted statistic from the j-th permuted data. In fact, using this set of permutation p-values and a valid multiple testing procedure (e.g., Holm's procedure), we are expected to control FWER as well. In addition to testing on permutation p-values, Li et al. (2012) proposed a more powerful approach that directly estimates FDP with the test statistics from resampled data.

The term *reproducibility* or *replicability* has recently attracted immerse attention in many scientific disciplines (e.g., Benjamini et al. (2009) and Baker (2016)). In a broader sense, reproducibility of scientific research implies that a different researcher can reproduce the experiment and reach the exactly same or highly similar conclusions (Fidler and Wilcox, 2018). In a narrower sense, it calls for reproducibility of computation (Peng, 2011). Given the same set of data and analysis pipeline, a different researcher should always obtain the same numerical results, such as list of significant discoveries. In the

context of this dissertation, we only focus on reproducibility in this narrower sense. Although the naïve decision rule that we mentioned above controls FDR, it can produce quite different lists of rejections (i.e., discoveries) when using a different permutation process (i.e., initiated by different random seeds), which are also different from the ideal decisions; in other words, it makes no effort to ensure reproducibility.

In literature, Gandy and Hahn (2014, 2016) were the only ones that have addressed the issue of non-reproducibility for multiple testing that requires resampling. To enhance reproducibility to a certain degree, they developed an algorithm based on resampling replicates to make decisions that aims to control the family-wise Monte Carlo error rate (MCER), an error rate closely related to reproducibility in terms of resampling process. Their originally proposed algorithm is sequential, giving a decision as *rejected*, *accepted*, or *undecided* to each test at each permutation step, and stopping sampling until there is no undecided test. Suppose that we test m (e.g., $m = 1000$) hypotheses $H_{1,0}, H_{2,0}, \cdots, H_{m,0}$. An MC error is defined as a disagreement with the ideal decisions among only tests that are determined as rejected or accepted:

$$V_i^{\mathrm{MC}} = \mathbb{I}(\widehat{D}_i = \textit{rejection}, D_i^* = \textit{acceptance}) + \mathbb{I}(\widehat{D}_i = \textit{acceptance}, D_i^* = \textit{rejection}), \quad (1.1)$$

where $D_1^*, D_2^*, \ldots, D_m^*$ and $\widehat{D}_1, \widehat{D}_2, \ldots, \widehat{D}_m$ are ideal decisions and decisions made by a certain resampling-based testing algorithm respectively. Let $V^{\mathrm{MC}} = \sum_{i=1}^m V_i^{\mathrm{MC}}$. Then the family-wise MCER is defined as $\Pr(V^{\mathrm{MC}} \geq 1)$. Controlling this error rate by α further implies that, two independent runs of this algorithm yield any disagreement among the decisions of tests that are determined as either rejected or accepted in both runs, with at most $1 - (1 - \alpha)^2 \approx 2\alpha$ chance. The undecided tests are those who require more permutation samples to make a decision between rejection and acceptance. There is no statement of reproducibility on undecided tests. Meanwhile, the methods of Gandy and Hahn (2014, 2016) control family-wise MCER at a pre-specified level even at an intermediate stage. Specifically, at each permutation step, they form a two-sided confidence interval for each underlying ideal p-value and apply the BH procedure to the collection of upper limits of intervals and lower limits separately. The tests whose both upper and lower limits have been rejected by the BH procedure will be determined as rejected, those

whose both limits have been accepted will be determined as accepted, and the remaining tests (if any) will be determined as undecided. Although this algorithm has rigorous error control, it is extremely conservative and tends to make zero rejection when the number of permutation samples is not very large.

When the number of permutation samples is fixed, in Chapter 3 we propose an algorithm that is more powerful than those algorithms under the framework of Gandy and Hahn (2014, 2016). The power is gained from two differences between our algorithm and the algorithms proposed by Gandy and Hahn. First, we control a different error rate from the family-wise MCER, which we call the family-wise type-I MCER, or family-wise MCER-I. A type-I MC error occurs when we reject a hypothesis which is accepted according to the ideal decisions (so that our decision on the hypothesis is non-reproducible):

$$V_i^{\text{MC-I}} = \mathbb{I}(\widehat{D}_i = \textit{rejection}, D_i^* = \textit{acceptance}), \tag{1.2}$$

where $\widehat{D}_1, \widehat{D}_2, \ldots, \widehat{D}_m$ are decisions made by our algorithm. Similarly, a type-II MC error occurs when $\widehat{D}_i = \textit{acceptance}$ but $D_i^* = \textit{rejection}$. Let $V^{\text{MC-I}} = \sum_{i=1}^m V_i^{\text{MC-I}}$. The family-wise MCER-I is the probability of observing one or more type-I MC errors, $\Pr(V^{\text{MC-I}} \geq 1)$. Thus our control of family-wise MCER-I is less stringent than the control of family-wise MCER. By controlling the family-wise MCER-I at level α, we are sure that the chance of any finding on our list being the result of MC error is less than α. Equivalently, the list of rejections by our method is a subset of the ideal list of rejections on ideal p-values with at least $(1-\alpha)$ probability. Although it is likely that we lose some rejections from the ideal list (to resolve it we need to control type-II MCER), this subset statement is surely a desirable property of resampling-based methods in a large-scale testing problem. All findings in our list would be rediscovered with at least $(1-\alpha)$ level of confidence, if a different researcher repeated this MC experiment using an infinite number of permutation samples. Due to this reason, we will first focus on controlling the type-I MCER. The second difference is that our algorithm is based on a fixed number of permutation samples. Gandy and Hahn (2016) suggested to use the Robbins-Lai confidence interval (Darling and Robbins, 1968; Robbins, 1970; Lai, 1976) to maintain exact coverage even when different tests stop after different number of permutations, due to early stopping for some

of the tests in the sequential procedure. However, the Robbins-Lai interval is generally too wide, particularly wider than the intervals using fixed permutation samples (Coe and Tamhane, 1993), and hence leads to loss of power. More details about this interval can be found at Section 1.3 and Section 4.2.1. This motivates us to develop a multiple testing procedure that controls MCER and is more powerful than the existing methods.

1.3 Sequential Resampling-Based Multiple Testing

Sequential algorithms provide an alternative solution to resampling-based multiple testing. We assume that there are m null hypotheses $H_{1,0}, H_{2,0}, \cdots, H_{m,0}$ and that the ideal p-values for m tests are $p_1^*, p_2^*, \cdots, p_m^*$ respectively. We rely on the standard BH procedure to control FDR. To be coherent with Section 1.2, a resampling replicate in either permutation or bootstrap test is called a permutation sample. With resampling-based multiple testing, it can be computationally intensive to evaluate significance accurately at each test. For instance, in genome wide association studies, we would need to sample 2.5×10^{13} permutation samples if we test $m = 500,000$ SNPs and run $n = 100m = 50,000,000$ permutations on each SNP. Sequential testing is a powerful tool to reduce the computational cost. In a sequential testing procedure, usually we specify a maximum number of permutations N_{max} (e.g., 10^6) in advance. Otherwise some resampling-based test possibly runs forever. Users can stop drawing further permutation samples if evidence suggests that the current approximated permutation p-value is sufficiently accurate. If for the i-th test the number of permutations that have been already sampled at the stopping time is n_i, the approximated permutation p-value $\widehat{p_i}$ can be given by either:

$$\frac{\sum_{k=1}^{n_i} \mathbb{I}(T_{i,k} > t_i)}{n_i} \text{ or } \frac{1 + \sum_{k=1}^{n_i} \mathbb{I}(T_{i,k} > t_i)}{n_i + 1}. \tag{1.3}$$

In both estimators above, t_i represents test statistic from the observed data, and $T_{i,k}$ represents the permuted statistic from the k-th permuted data.

For a single resampling-based test, Besag and Clifford (1991) proposed a simple stopping rule which defines n_i to be the smallest integer such that $\sum_{k=1}^{n_i} \mathbb{I}(T_{i,k} > t_i) \geq n_0$

where n_0 is a fixed positive integer that ensures accurate approximation (Besag and Clifford (1991) suggested $n_0 = 10$ or 20), or $n_i = N_{max}$ if such an integer does not exist. It can be shown that given this sequential decision rule the estimator

$$\widehat{p}_i^{(BC)} = \begin{cases} n_0/n_i & \text{if } n_i < N_{max} \\ \frac{1+\sum_{k=1}^{n_i} \mathbb{I}(T_{i,k} > t_i)}{n_i+1} & \text{if } n_i = N_{max}, \end{cases}$$

is more conservative than U[0, 1] under the i-th null, meaning that $\Pr(\widehat{p}_i^{(BC)} \leq u | H_{i,0}) \leq u$. Sandve et al. (2011) considered a sequential multiple testing procedure in which this stopping rule is applied on each individual test and permutation p-values are estimated accordingly. The Sandve's procedure controls the FDR that is calculated by truth (a test being null or non-null). However, Gandy and Hahn (2014) reported that the Sandve's procedure may severely suffer from MC errors: many rejected or accepted tests are actually decided randomly; with the same data and testing algorithm, decisions change easily once we choose a different random seed. Several papers discussed MC errors in the context of resampling-based test, but were limited to testing single hypothesis, including Fay and Follmann (2002), and Kim (2010). Guo and Peddada (2008) and Jiang and Salzman (2012) considered a different question and developed sequential multiple testing algorithms which produce the same set of decisions as the non-sequential problem with a fixed number of permutation samples. Their procedures can be shown to control FDR up to a small error term. Both methods cannot provide theoretical guarantee on MC error rate control.

In the aspect of applied statistics, a procedure that cannot replicate its results is usually unsatisfactory. Similar to the direction pursued by Gandy and Hahn, our goal is to develop a sequential multiple testing procedure that enjoys high reproducibility on the set of rejected or accepted tests. The problem of controlling MCER in sequential resampling-based tests was first introduced by Gandy (2009). Gandy and Hahn (2014, 2016) further extended this concept to multiple testing problems. Gandy and Hahn (2016) developed a general statistical framework to control the family-wise MCER in resampling-based multiple testing. We sketch the main steps of the sequential testing algorithm proposed by Gandy and Hahn (2016). First, without considering any stopping

rule, simultaneous confidence intervals $(p_1^*, p_2^*, \cdots, p_m^*)$ of ideal p-values are created by the Bonferroni correction. Suppose we want to control the family-wise MCER by α. The confidence interval for each p^* is a $(1-\alpha/m)$ level Robbins-Lai (sequential) confidence interval, which is denoted by $\mathcal{I}_i = (p_i^l, p_i^u)$. The Robbins-Lai confidence interval can be calculated by considering the roots of the equation below. We use the i-th test as an example:

$$(k+1)\binom{k}{X_{i,k}} p_i^{X_{i,k}} (1-p_i)^{k-X_{i,k}} = \frac{\alpha}{m}. \tag{1.4}$$

Here $\alpha \in (0,1)$ and $X_{i,k} = \sum_{k'=1}^{k} \mathbb{I}(T_{i,k'} > t_i)$ means the number of permuted statistics that are more extreme than the observed statistic at the time of collecting exactly k permutation samples. Let $\mathcal{I}_{i,k}$ be the set

$$\left\{ p_i \in (0,1) \;\middle|\; (k+1)\binom{k}{X_{i,k}} p_i^{X_{i,k}} (1-p_i)^{k-X_{i,k}} > \frac{\alpha}{m} \right\}. \tag{1.5}$$

If $X_{i,k} = 0$, there is only one root $p_{i,k}^u$ in (1.4) and $\mathcal{I}_{i,k} = (0, p_{i,k}^u)$; if $X_{i,k} = k$, there is only one root $p_{i,k}^l$, and $\mathcal{I}_{i,k} = (p_{i,k}^l, 1)$; if $0 < X_{i,k} < k$, there are two distinct roots $p_{i,k}^l < p_{i,k}^u$, and $\mathcal{I}_{i,k} = (p_{i,k}^l, p_{i,k}^u)$. According to Lai (1976), the coverage of the interval $\mathcal{I}_i = \cap_{k=1}^{n_i} \mathcal{I}_{i,k}$ is always above $(1-\alpha/m)$ for any stopping time n_i. After we obtain the simultaneous confidence intervals, we have

$$\Pr(p_1^* \in \mathcal{I}_1, p_2^* \in \mathcal{I}_2, \cdots, p_m^* \in \mathcal{I}_m) \geq 1-\alpha$$

which means that the ideal p-values being included by the rectangular region $\times_{i=1}^{m} \mathcal{I}_i = \mathcal{I}_1 \times \mathcal{I}_2 \times \cdots \times \mathcal{I}_m$ with at least $(1-\alpha)$ probability. The i-th test is decided to be *rejected* if both upper limits and lower limits of $\mathcal{I}_i$ are rejected by standard BH, or *accepted* if both upper limits and lower limits of $\mathcal{I}_i$ are accepted by standard BH. If both upper limits and lower limits agree with the decision, any point inside $\times_{i=1}^{m} \mathcal{I}_i$ should yield the same decision. This is why the probability of occurring none MC error is at least above the simultaneous coverage probability $(1-\alpha)$. Or equivalently the family-wise MCER is at most α. After a decision being made, we will not further generate permutation samples for that hypothesis. Usually at the middle and late stage of permutation testing, we only need to draw samples on a small number of hypotheses. This is a reason why sequential

resampling-based testing is quite computationally efficient. Details of this algorithm are presented in Algorithm 1.

Initialize $\widehat{D}_1, \widehat{D}_1, \cdots, \widehat{D}_m$ to be *undecided*.
Initialize $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ to be $(0, 1)$.
▷ Permutation begins
for *k in* $1 : \mathrm{N}_{\max}$ **do**
  for *i in* $1 : m$ **do**
    if $\widehat{D}_i = \textit{undecided}$ **then**
      Update $\mathcal{I}_i = \mathcal{I}_i \cap \mathcal{I}_{i,k}$ where $\mathcal{I}_{i,k}$ is defined in (1.5).
    end
  end
  Apply standard BH on the upper limits of intervals $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ and decisions are $d_1^u, d_2^u, \cdots, d_m^u$;
  Apply standard BH on the lower limits of intervals $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ and decisions are $d_1^l, d_2^l, \cdots, d_m^l$.
  for *i in* $1 : m$ **do**
    if $d_i^u = d_i^l = \textit{acceptance}$ **then**
      Update $\widehat{D}_i =$*acceptance*.
    end
    if $d_i^u = d_i^l = \textit{rejection}$ **then**
      Update $\widehat{D}_i =$*rejection*.
    end
  end
end
▷ Permutation ends
return $(\widehat{D}_1, \widehat{D}_1, \cdots, \widehat{D}_m)$.

Algorithm 1: The sequential multiple resampling-based tests proposed by Gandy and Hahn (2016)

Although the methods can be useful under many different scenarios, low statistical power is usually a concern. Many tests cannot be decided even with a very large number of permutations. One factor that may cause the loss of power is that they used the Bonferroni correction to construct simultaneous confidence intervals for estimating ideal p-values. It is known that the Bonferroni correction is extremely conservative in most applications, especially when the number of tests is so large. A remedy to this problem is replacing the Bonferroni correction with step-wise procedures which are known to bring more power in multiple testing literature. The closure principle (Marcus et al., 1976) and partition principle (Finner and Strassburger, 2002) are the two fundamental principles related to general step-wise procedures. A series of multiple testing methods

that control FWER, say Holm (1979); Bauer et al. (1998); Romano and Wolf (2005); Hommel et al. (2007); Bretz et al. (2009); Dobriban (2018) can be developed with the closure principle. The partition principle is primarily useful for constructing step-wise simultaneous confidence intervals. Based on the partition principle, it can be shown that the one-sided simultaneous confidence intervals that are compatible with the Holm's procedure always exist (Strassburger and Bretz, 2008).

We discussed two types of error rates, family-wise MCER-I and family-wise MCER in section 1.2. Compared to family-wise MCER-I, family-wise MCER seems a more reasonable error rate for sequential multiple testing. Unlike the scenario where we have a fixed budget of generating permutation samples, in a sequential problem we can promise to assign the majority of tests with either rejection or acceptance decisions as long as N_{max} is a large number.

Chapter 2

A Bottom-up Approach to Testing Hypotheses That Have a Branching Tree Dependence Structure, with False Assignment Rate Control

2.1 Introduction

In this chapter we develop a bottom-up test to identify associated taxa in microbiome data. In Section 2.2.2 we introduce a modified null hypothesis for bottom-up testing that adjusts for selection decisions at lower levels of tree. We further develop an error criterion we call the false assignment rate (FAR) that corresponds to this modified null hypothesis. From Section 2.2.3 to 2.2.5, we propose an algorithm for assessing the significance of association between taxonomic units (or, more generally, nodes in the tree) and a trait under study that controls the FAR. In these three sections we introduce a procedure for unweighted testing, a procedure for weighted testing, and a procedure that handles incomplete trees. Section 2.2.6 is an extension to controlling error rates separately. In Section 2.3, we compare our proposed methods with other existing methods using simulated data, and show that the FAR can approximate the FDR under the conjunction null hypothesis. In Section 2.4, we apply our new methods to data on the human gut microbiome from a study of inflammatory bowel disease (IBD), and detect driver taxa that are associated with ulcerative colitis (UC).

2.2 Methods

2.2.1 Preliminaries

The hypotheses we test form the nodes of a branching tree; here, we review the terminology we use. The *root* node is the "top" of the tree (in Figure 2.1(a), $N_{4,1}$ is the root node). For any two nodes that are directly connected, the node closest to (furthest from) the root is the *parent* (*child*) node. The set of child nodes of a parent are its *offspring*. A node is an *inner node* if it has at least one child node; otherwise it is a *leaf* node. The *ancestors* of a node are all the nodes traversed in a path from that node to the root. The *descendents* of a node are all nodes having that node as an ancestor. A *subtree* is a tree rooted at an inner node of the full tree, comprised of the subtree root and all its descendants. For example, in Figure 1(a), the tree rooted at $N_{3,1}$ that includes inner nodes $N_{3,1}$, $N_{2,1}$, $N_{2,2}$, and leaf nodes $N_{1,1}$, $N_{1,2}$, $N_{1,3}$, $N_{1,4}$ is a subtree of the full

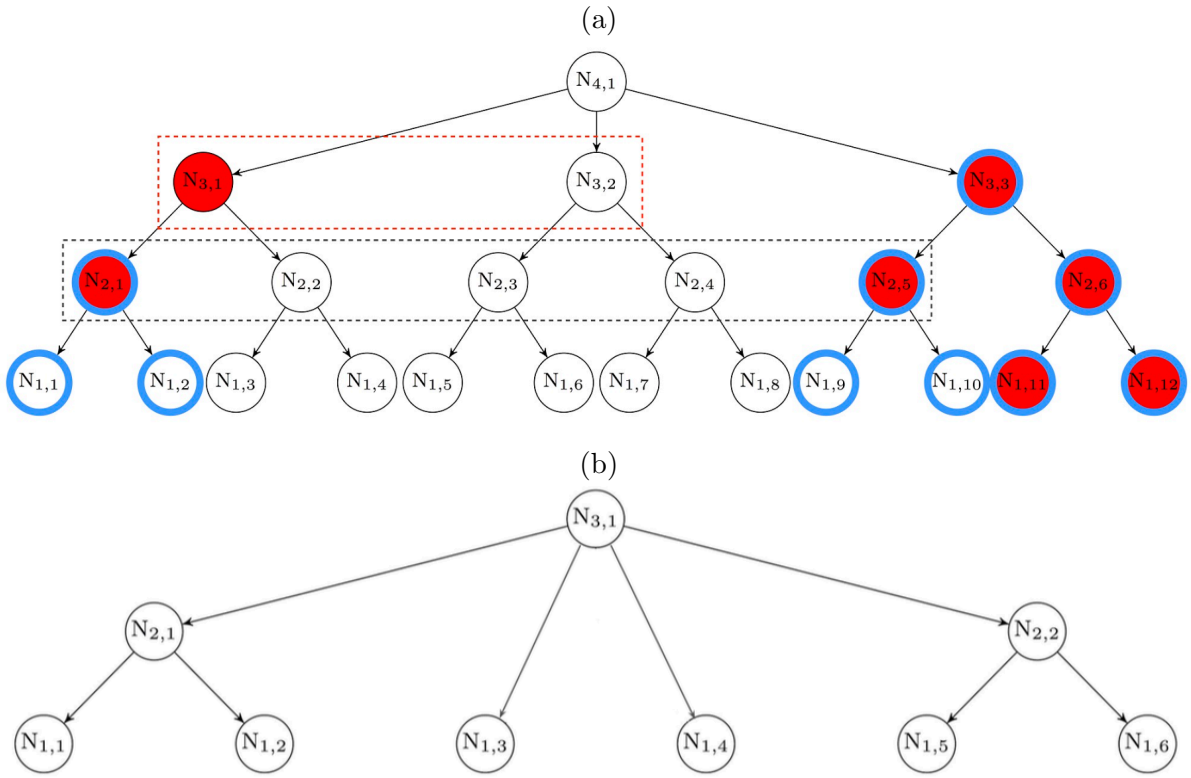

Figure 2.1: (a) A hypothetical example of a set of hypotheses having a tree-structured relationship. Nodes are labeled by *level* (first subscript) and then numbered within level (second subscript). Nodes highlighted with blue circles are truly associated. A node colored red indicates it is detected (declared to be associated with the trait of interest by a testing method). With the bottom-up methods, all nodes at the bottom level are tested at level 1, nodes inside the black box are tested at level 2, and nodes inside the red box are tested at level 3. (b) A hypothetical example illustrating a set of hypotheses having a dependence structure corresponding to an incomplete tree. In this example, it makes scientific sense to assign nodes $N_{1,3}$ and $N_{1,4}$ to level 1 even though they have a different depth than the other leaf nodes. For example, these two nodes could correspond to OTUs that are missing a species assignment but share a genus with the other leaf nodes.

tree. The *depth* of a node is the number of edges between that node and the root node. Because each node corresponds to a hypothesis, we will sometimes refer to testing a node as shorthand for testing the hypothesis at that node.

We use the term *level* to describe sets of nodes that will be tested together. In the simple case such as in Figure 2.1(a), nodes that have the same depth are assigned to the same level; we call such a tree *complete*. For *incomplete* trees such as that shown in Figure 2.1(b), level is assigned by the investigator and does not necessarily correspond to depth. For example, in a phylogenetic (taxonomic) tree, level typically corresponds to the taxonomic rank (species, genus, etc.); a phylogenetic tree is then incomplete when the leaf nodes (OTUs) have missing assignment below a certain level. The tree shown in

Figure 2.1(b) could be an example of this, where level 1 corresponds to OTUs and where OTUs $N_{1,1}, N_{1,2}, N_{1,5}$, and $N_{1,6}$ have genus *and* species assignments but OTUs $N_{1,3}$ and $N_{1,4}$ are *only* assigned at the genus level. We will sometimes refer to a node at which we have rejected the null hypothesis as a *detected* node, and a detected node is a *driver* node if none of its ancestors are detected. We further assume that p-values for the association tests at all leaf nodes are available. The algorithm we describe here gives p-values for all internal nodes as needed.

There are two ways to imagine calculating p-values for inner nodes in a bottom-up testing algorithm for tree-structured hypotheses. In the first approach, p-values for inner nodes are determined entirely from the p-values of their offspring (and hence, are determined by the p-values at leaf nodes). In the second approach, p-values for inner nodes are calculated by applying a test statistic to pooled data (Tang et al., 2017). The second approach may be problematic as it may be hard to determine the null distribution of the pooled data, given decisions about p-values at lower levels (e.g., removing data from nodes having p-values less than a threshold from pooling for testing the modified null defined in Section 2.2.3). Also, it is hard to know how the conjunction null hypothesis could be tested using pooled data. In addition, pooling data may result in effects being cancelled out if some offspring nodes are protective while others increase risk. For these reasons, we seek an algorithm that operates entirely on the p-values of the leaf nodes.

2.2.2 Bottom-up Testing

The goals of our inference are to find nodes in the tree (e.g., taxa for the microbiome example) which are associated with a trait of interest. We wish to avoid declaring a node to be associated just because a few offspring nodes are strongly associated; thus we restrict claims of association to nodes in which a large number of offspring are associated. For this goal the global null hypothesis, which specifies a node is associated if *even one* offspring node is associated, is not appropriate. Directly testing the conjunction null hypothesis that *not all* offspring of a node are associated (Price and Friston, 1997) is known to be conservative in many situations, as the p-value is determined by selecting the *largest* p-value from the p-values at each offspring, and then comparing the selected

p-value to the uniform distribution on $[0, 1]$ (Friston et al., 2005). However, it does easily lead to a bottom-up procedure; after propagating the largest p-value from offspring nodes to their parent nodes, nodes are then detected using the standard BH (Benjamini and Hochberg, 1995) procedure. We report results from this procedure even though we do not recommend it, due to its low power.

To develop a bottom-up procedure that avoids the low power of the rigorous test of the conjunction null hypothesis, but still finds taxa for which most offspring are associated, we use the insight that, if an offspring node has already been found to be associated, including it in our test for the parent adds no new information. This insight is reinforced by the observation that, in an omnibus test of association for all offspring, a strong association from already-detected children may lead to the parent node being detected even if most other child nodes are truly null. Thus, as a surrogate for the conjunction null hypothesis, we propose to test a modified null that, among the offspring *that have not been previously detected at a lower level*, none are associated, against the alternative that some previously-undetected offspring are associated. We then combine the p-values of these previously-undetected nodes to form a test statistic for the parent node, using a small modification of Stouffer's Z test. Nodes for which all offspring are already detected are not tested, but automatically detected and require special handling as described in Section 2.2.4.

To illustrate these issues, consider the hypothetical example in Figure 2.1(a). Nodes highlighted with blue circles are truly associated; $N_{2,1}$ and $N_{3,3}$ are driver taxa. Although $N_{3,1}$ is associated under the global null hypothesis because $N_{1,1}$ and $N_{1,2}$ are its descendants, we would prefer to conclude that $N_{2,1}$ rather than $N_{3,1}$ explains the association signal among the descendants of $N_{3,1}$ because $N_{3,1}$ has descendants that are not truly associated. This is achieved by using the modified null hypothesis, because $N_{3,1}$ is *not* associated under the modified null hypothesis, because $N_{2,2}$ is not associated.

The error rate of any testing algorithm depends on the null hypothesis used to ascertain the true association status of each node. Thus, we distinguish between the FDR, for which we use the global null hypothesis to determine true association status; the false assignment rate (FAR), for which we use the modified null to determine true association

status; and the FDRc, for which we use the conjunction null to determine true association status. For all three error rates, a false discovery means that a node was detected (i.e., found to be associated) that is in fact not associated, under the appropriate null. We use the term FAR rather than modified FDR because the "assignment" of nodes as being associated or not is influenced by decisions made at lower levels, and reserve the term "discovery" for situations where decisions are based on a test statistic that is calculated for each node without regard to decisions at other nodes. Thus, our procedure must test all nodes at the lowest level, then the next lowest level and so on. Suppose the tree has L levels, and let n_l, $l = 1, 2, \ldots, L$, be the number of nodes on level l and $N_{l,j}$, $j = 1, 2, \ldots, n_l$, denote the j^{th} node on level l. We then define

$$R_{l,j} = \mathbb{I}\left(\text{node } N_{l,j} \textit{ is detected}\right),$$

where $\mathbb{I}(.)$ is the indicator function. We note that if node $N_{l,j}$ is not tested, then $R_{l,j} = 0$ by default. Let $\mathcal{U}_{l,j}$ denote the set of undetected offspring of the j^{th} node on level $l = 2, \ldots, L$; note that $\mathcal{U}_{1,j} = \emptyset$. We next define $V^x_{l,j}$ to indicate a false assignment was made under null hypothesis x, where $x = g$ for the global null hypothesis, $x = m$ for the modified null hypothesis, and $x = c$ for the conjunction null hypothesis. Thus,

$$V^m_{l,j} = \mathbb{I}\left(R_{l,j} = 1 \text{ but the } \textit{modified} \text{ null hypothesis given } \mathcal{U}_{l,j} \text{ at node } N_{l,j} \text{ is true}\right).$$

Similarly, for the global and conjunction null hypotheses, we define $V^g_{l,j}$ and $V^c_{l,j}$ as

$$V^x_{l,j} = \mathbb{I}\left(R_{l,j} = 1 \text{ but null hypothesis } x \text{ at node } N_{l,j} \text{ is true}\right), \; x = c, g.$$

The error rate under each null hypothesis is given by

$$\mathbb{E}\left[\frac{\sum_{l=1}^{L}\sum_{j=1}^{n_l} V^x_{l,j}}{\left(\sum_{l=1}^{L}\sum_{j=1}^{n_l} R_{l,j}\right) \bigvee 1}\right].$$

If the global null hypothesis at $N_{l,j}$ is true, then the modified null hypothesis at $N_{l,j}$ must also be true, which in turn implies the conjunction null hypothesis at $N_{l,j}$ is true. Thus, $V^g_{l,j} \leq V^m_{l,j} \leq V^c_{l,j}$ holds for all nodes. Thus, the three error rates FDR, FAR and FDRc

defined above, are related by

$$\text{FDR} \leq \text{FAR} \leq \text{FDRc}.$$

This implies that controlling FAR is a more stringent criterion than controlling FDR, and so a testing procedure that controls the FAR will automatically control the FDR. However, controlling FAR does not guarantee control of FDRc. Nevertheless, to the extent that the test used to combine the p-values of previously-undetected nodes is powerful when non-null effects appear in most or all individual tests, we can expect controlling FAR should be similar to controlling FDRc. We return to this issue in section 2.2.3, where this reasoning leads us to advocate use of Stouffers Z-score over Fisher's method, when testing the modified null hypothesis.

The modified null hypothesis we test has another important implication: a test at one level may decide hypotheses at one or more higher levels. This occurs when a node has no undetected offspring, i.e. $\mathcal{U}_{l,j} = \emptyset$. For example, since both offspring of $N_{2,6}$ in Figure 2.1(a) are detected, we should immediately conclude that $N_{2,6}$ is associated. Similarly, having already detected $N_{2,6}$, if we determine that $N_{2,5}$ is associated, then $N_{3,3}$ would have no undetected offspring and should be determined to be associated as well. We present two approaches to account for the effect that this multiplicity has on the FAR. In the first approach, we do not allow this propagation, and instead consider that the test of the last undetected offspring of a parent node is in fact a test of the parent node (or, in general, a test of the highest node decided by this single test). So, for example, in Figure 2.1(a), if $N_{2,6}$ had already been detected, rejecting the modified null at $N_{2,5}$ would add $N_{3,3}$ to the list of detected nodes, but not $N_{2,5}$. In this way, each hypothesis we test results in a single addition to the list of detected nodes. Although this solution is unsatisfactory in many ways, it leads to a simpler procedure that serves as a useful intermediate result in developing our recommended approach (which we present in Section 2.2.4).

2.2.3 Testing to Control FAR: Unweighted Proposal

We now construct a testing procedure that tests the nodes in the tree level by level, starting at level $l = 1$. For each level $l = 1, \ldots, L$, our testing procedures consist of two elements: a set of thresholds to determine which nodes are detected at level l, and a way of aggregating the p-values from the undetected nodes at level l to give p-values for the (parent) nodes at level $l+1$ for those nodes at level $l+1$ that have undetected offspring. Our goal is to control the error rate so that the FAR $\leq q$. In analogy with the concept of alpha spending in interim analysis (Demets and Lan, 1994), we allocate to each level l a target level q_l ($l = 1, 2, \ldots, L$) chosen so that $\sum_{l=1}^{L} q_l = q$. We note here that we do not guarantee the FAR *at each level* is controlled at level q_l, just that the *overall* FAR is controlled at level q. Although it would be interesting to develop an optimal strategy for choosing the q_ls, we choose $q_l = qn_l/n$, where n_l is the number of nodes at level l and $n = \sum_{l=1}^{L} n_l$ is the number of nodes in the tree.

We first consider how to assign p-value thresholds to control FAR. Recall that for $l > 1$, tests at a lower level may have already resulted in detection of some of the nodes at level l. Suppose that, of the n_l total nodes at level l, there are n_l^* nodes that have at least one child node that has not been detected. Without loss of generality, assume the p-values for each node at level l, $p_{l,1}, p_{l,2}, \ldots, p_{l,n_l^*}$, have been sorted in ascending order, and let the sorted values be denoted by $p_{l,(1)} \leq p_{l,(2)} \leq \cdots \leq p_{l,(n_l^*)}$. Let d_l^* denote the (as yet unknown) number of nodes detected at level l. We seek a set of ascending thresholds $\alpha_{l,1} \leq \alpha_{l,2} \leq \cdots \leq \alpha_{l,n_l^*}$ by which we reject the modified null hypothesis at $d_l^* > 0$ nodes (corresponding to $p_{l,(1)}, \ldots, p_{l,(d_l^*)}$) if $p_{l,(1)} \leq \alpha_{l,1}, p_{l,(2)} \leq \alpha_{l,2}, \ldots, p_{l,(d_l^*)} \leq \alpha_{l,d_l^*}$ but $p_{l,(d_l^*+1)} > \alpha_{l,d_l^*+1}$; we accept the modified null hypothesis at all nodes in level l if $p_{l,(1)} > \alpha_{l,1}$ in which case we take $d_l^* = 0$, or reject the modified null hypotheses at all nodes in level l if $p_{l,(1)} \leq \alpha_{l,1}, p_{l,(2)} \leq \alpha_{l,2}, \ldots, p_{l,(n_l^*)} \leq \alpha_{l,n_l^*}$ in which case we take $d_l^* = n_l^*$. We adopt the thresholds $\{\alpha_{l,j}\}$ given by

$$\frac{\alpha_{l,j}}{1-\alpha_{l,j}} = \left(\frac{D_{l-1}+j}{n_l^*-j+1} \times q_l\right) \bigwedge \frac{\tau_0}{1-\tau_0}, \tag{2.1}$$

where $D_{l-1} = \sum_{l'=1}^{l-1} d_{l'}^*$ for $l \geq 2$ is the cumulative number of detection made up to and

including the $(l-1)$th level, $D_0 = 0$, and τ_0 is a pre-specified constant to prevent nodes with large p-values from being detected if a large number (say, m) of null hypotheses can be easily rejected because of very low p-values, in which case $q \times m$ nodes with large p-values can be said to be detected, while still controlling the overall error rate at level q; we set $\tau_0 = 0.5$ in this article, which follows the same choice in Storey (2002). At each level, the thresholds (2.1) are a variant of the thresholds in the step-down test proposed by Gavrilov et al. (2009), which have been used to control FDR in some applications as they have been shown to be more powerful than the standard BH procedure. Theorem 2.1 asserts that our bottom-up procedure with thresholds (2.1) control the FAR at q.

Theorem 2.1. Assume the three conditions hold: (C1) nodes on the same level have the same depth; (C2) p-values for null nodes follow the uniform distribution U$[0,1]$; (C3) at each level, the p-value for a null node is independent of the p-values at all other nodes. Then the bottom-up procedure with thresholds (2.1) ensures that the FAR $\leq q$.

The proof of this theorem is provided in Appendix A.1. Condition (C1) assumes that nodes on the same level have the same depth, and will be relaxed in Section 2.2.5. Conditions (C2) and (C3) can be satisfied by our proposal below for obtaining p-values for parent nodes.

We now consider how to aggregate the p-values from level l that correspond to the undetected offspring of a node at level $l+1$. Note that each undetected node at level l is a member of exactly one of the sets $\mathcal{U}_{l+1,j}$, $j = 1, \ldots, n^*_{l+1}$, the collections of the undetected offspring of nodes at level $l+1$. Thus, for the j^{th} node at level $l+1$, we pool information from nodes in $\mathcal{U}_{l+1,j}$. Note that the p-values of the undetected nodes at level l necessarily exceed the threshold α_{l,d^*_l+1}, and are hence not uniformly distributed on the interval $[0,1]$. However, since this is the only restriction on these p-values, it follows that, under either the global or modified null hypothesis, the p-values for nodes that were not detected at level l are uniformly distributed on the interval $[\alpha_{l,d^*_l+1}, 1]$; equivalently, adjusted p-values $p'_{l,k} = (p_{l,k} - \alpha_{l,d^*_l+1})/(1 - \alpha_{l,d^*_l+1})$ are uniformly distributed on the interval $[0,1]$. Thus,

we form Stouffer's Z score:

$$Z_{l+1,j} = \frac{1}{\sqrt{|\mathcal{U}_{l+1,j}|}} \sum_{k \in \mathcal{U}_{l+1,j}} \Phi^{-1}(1 - p'_{l,k}), \tag{2.2}$$

where Φ is the standard normal cumulative distribution function and $|\mathcal{U}_{l+1,j}|$ represents the cardinality of $\mathcal{U}_{l+1,j}$. The $Z_{l+1,j}$ calculated using the undetected null nodes in $\mathcal{U}_{l+1,j}$ follows a standard normal distribution $N(0,1)$ under the modified null conditional on the pattern of detection at level l. Thus, $p_{l+1,j} = 1 - \Phi(Z_{l+1,j})$; in addition, $p_{l+1,j}$ and $p_{l+1,j'}$ are independent since, on any tree, $\mathcal{U}_{l+1,j} \cap \mathcal{U}_{l+1,j'} = \emptyset$.

We use Stouffer's Z-score as it is known to be powerful when small or moderate non-null effects appear in the majority of individual tests as opposed to Fisher's method, which is additionally powerful when only a few large non-null effects are present (Loughin, 2004). As a result, using Stouffer's Z-score when controlling the FAR gives a better control of FDRc than using Fisher's method (the results based on Fisher's method not shown).

To summarize our procedure, we start at level $l = 1$ with the p-values at the leaf nodes (which are given). We determine which are detected and which are not detected using thresholds $\alpha_{1,j}$ calculated using (2.1). For any nodes at level $l = 2$ that have undetected offspring (i.e., for which $\mathcal{U}_{2,j} \neq \emptyset$), we then aggregate these undetected p-values into a Z score using (2.2) and convert the value of this statistic into a p-value, which then serves as the p-value for the (parent) nodes on level $l = 2$. We continue in this manner until we reach the root node of the tree. For this simplified approach, our list of detected nodes consists of each node that was detected at level l for nodes that did not result in multiple detection for a single test, or the highest node detected for those nodes whose detection resulted in nodes at a higher level also being detected.

2.2.4 Testing to Control FAR: Weighted Procedure

The testing procedure described in Section 2.2.3 may be unsatisfactory to many users because, when a test at a single node results in detection of multiple nodes, only one can be included on the list of discoveries if we wish to control FAR. Thus, we consider a modification of this algorithm which allows all the detected nodes to be considered

discoveries.

The difficulty with including all discoveries made by the approach in Section 2.2.3 is that an incorrect decision on a single node can result in multiple false discoveries. To resolve this, we introduce weights $\omega_{l,j}$ that count the number of detections that could arise when testing node $N_{l,j}$ under the modified null hypothesis (i.e., conditional on the detections at lower levels). Consider the example shown in Figure 2.1(a). Assume at level 1, nodes $N_{1,11}$ and $N_{1,12}$ have been determined to be significant. Then, at level 2, the remaining nodes to be tested are $N_{2,1}$–$N_{2,5}$ (inside the black box). We consider testing nodes at each level in ascending order of p-values and assume $p_{2,1} < p_{2,2} < p_{2,3} < p_{2,4} < p_{2,5}$. Rejecting the modified null hypothesis at $N_{2,1}$ only detects $N_{2,1}$; rejecting the null at $N_{2,2}$ will detect $N_{2,2}$ and $N_{3,1}$ (2 nodes); rejecting the null at $N_{2,3}$ will detect $N_{2,3}$; rejecting the null at $N_{2,4}$ will detect $N_{2,4}$ and $N_{3,2}$ (2 nodes); rejecting the null at $N_{2,5}$ will detect $N_{2,5}$, $N_{3,3}$, and $N_{4,1}$ (3 nodes). Thus, for this ordering, we define the weights $(\omega_{2,1}, \omega_{2,2}, \omega_{2,3}, \omega_{2,4}, \omega_{2,5}) = (1, 2, 1, 2, 3)$. Now, suppose instead that the p-values were ordered as $p_{2,5} < p_{2,1} < p_{2,2} < p_{2,3} < p_{2,4}$; then the weights would be $(\omega_{2,5}, \omega_{2,1}, \omega_{2,2}, \omega_{2,3}, \omega_{2,4}) = (2, 1, 2, 1, 3)$. Although these weights are different, the *sorted* weights $(\omega_{2,(1)}, \omega_{2,(2)}, \omega_{2,(3)}, \omega_{2,(4)}, \omega_{2,(5)}) = (1, 1, 2, 2, 3)$ are the same. It is easy to verify that the same set of *sorted weights* will be obtained with any other ordering of nodes. Thus, the (unsorted) weights $(\omega_{l,1}, \omega_{l,2}, \ldots, \omega_{l,n_l^*})$ depend on the p-values at level l and are thus random (even conditional on the detection events below level l). However, for complete trees (i.e., under Condition C1), we show in Appendix A.2 that the sorted weights $\omega_{l,(1)} \leq \omega_{l,(2)} \leq \cdots \leq \omega_{l,(n_l^*)}$ are unique regardless of the ordering of p-values.

Using the weights just defined, the FAR we wish to control becomes

$$\text{FAR} = \mathbb{E}\left[\frac{\sum_{l=1}^{L}\sum_{j=1}^{n_l^*}\omega_{l,j}V_{l,j}^m}{\left(\sum_{l=1}^{L}\sum_{j=1}^{n_l^*}\omega_{l,j}R_{l,j}\right)\bigvee 1}\right]. \tag{2.3}$$

We modify the thresholds in (2.1) by replacing the count j with $\sum_{k=1}^{j}\omega_{l,(k)}$ and $n_l^* - j + 1$ with $\sum_{k=j}^{n_l^*}\omega_{l,(k)}$ to give:

$$\frac{\alpha_{l,j}}{1-\alpha_{l,j}} = \left(\frac{D_{l-1} + \sum_{k=1}^{j}\omega_{l,(k)}}{\sum_{k=j}^{n_l^*}\omega_{l,(k)}} \times q_l\right) \bigwedge \frac{\tau_0}{1-\tau_0}. \tag{2.4}$$

Theorem 2.2 (proved in Appendix A.2) asserts that the bottom-up procedure with thresholds (2.4) controls FAR (2.3) at value $\leq q$.

Theorem 2.2. Under Conditions (C1), (C2) and (C3) in Theorem 2.1, the bottom-up procedure with thresholds (2.4) ensures that the value of the FAR given in equation (2.3) is $\leq q$.

2.2.5 Bottom-up Testing on Incomplete Trees

In Sections 2.2.3–2.2.4, we only considered complete trees where nodes on the same level all have the same depth. Here we consider more general trees where depth and level do not coincide. For example, in the tree from Figure 2.1(b), nodes $N_{1,3}$ and $N_{1,4}$ have different depth from the other leaf nodes, although they are all on the same level. In the microbiome example, this would occur whenever some of the lower taxonomic ranks (e.g., species and genus) of an OTU are not known. One possible solution is to fill in the missing levels by assigning each such OTU its own (unknown) species and (unknown) genus. This is unsatisfactory both scientifically, as we then assert that the "correct" species and genus for this OTU is different from any other OTU, and statistically, as the p-value for the species and genus level tests are necessarily identical to the p-value for the OTU. Although this may seem similar to the situation addressed in Section 2.2.4 where a single test could determine multiple hypotheses, it is actually different in that there is no additional information gained at the species or genus level in this example. A better alternative is to place each leaf node at the level just below the nearest inner nodes. However, this strategy can still be unsatisfactory for applications such as the microbiome, as it is scientifically questionable to treat some OTUs at the same level as higher taxa such as families. As a result, we describe here how our approach can be extended to incomplete trees.

For an incomplete tree, the sorted weights $(\omega_{l,(1)}, \omega_{l,(2)}, \ldots, \omega_{l,(n_l^*)})$ are no longer unique. In Figure 2.1(b), when $N_{1,6}$ has the largest p-value among all leaf nodes, the sorted weights at level 1 are $(1, 1, 1, 1, 2, 3)$; if $N_{1,4}$ has the largest p-value, the sorted weights are $(1, 1, 1, 2, 2, 2)$. To account for this ambiguity, we seek a single set of sorted weights that

will control FAR for any possible ordering of p-values. For the two sets of weights just considered, note that the cumulative sums of sorted weights $\sum_{k=1}^{j} \omega_{l,(k)}$ for the first set, given by $(1, 2, 3, 4, 6, 9)$ are all less than or equal to the cumulative sums of the sorted weights of the second set, given by $(1, 2, 3, 5, 7, 9)$. Thus, if we were to use the first set of ordered weights in (4), the thresholds $\alpha_{l,j}$ would be smaller than the thresholds calculated using the second set of sorted weights. In Appendix A.3, we show how to find a unique set of sorted weights $\widetilde{\omega}_{l,(1)} \leq \cdots \leq \widetilde{\omega}_{l,(n_l^*)}$ for level l that correspond to weights obtained by some ordering of p-values and that satisfy the inequalities

$$\sum_{k=1}^{j} \widetilde{\omega}_{l,(k)} \leq \sum_{k=1}^{j} \omega_{l,(k)}, \quad j = 1, \ldots, n_l^*,$$

for all possible sets of sorted weights $(\omega_{l,(1)}, \ldots, \omega_{l,(n_l^*)})$ induced by different orderings of p-values. We then adopt thresholds calculated using $\{\widetilde{\omega}_{l,(k)}\}$, given by

$$\frac{\alpha_{l,j}}{1 - \alpha_{l,j}} = \left(\frac{D_{l-1} + \sum_{k=1}^{j} \widetilde{\omega}_{l,(k)}}{\sum_{k=j}^{n_l^*} \widetilde{\omega}_{l,(k)}} \times q_l \right) \bigwedge \frac{\tau_0}{1 - \tau_0}. \tag{2.5}$$

Because these thresholds are the most stringent among those based on any possible weights $(\omega_{l,(1)}, \ldots, \omega_{l,(n_l^*)})$, we call them the least favorable weights. Theorem 2.3 ensures control of the FAR using the least favorable weights.

Theorem 2.3. Under Conditions (C2) and (C3) in Theorem 2.1, the bottom-up procedure with thresholds (2.5) ensures the FAR defined in (2.3) is $\leq q$.

The proof of Theorem 2.3 can be found in Appendix A.4. When a tree is complete, the least favorable weights reduce to the unique sorted weights regardless of the ordering of p-values. Thus the testing procedure presented here encompasses the one presented in Section 2.2.4 as a special case.

2.2.6 Bottom-up Testing with Separate FAR Control

The testing procedures we have described so far assume that we wish to detect nodes at all levels of the tree while controlling the overall FAR at some level q. In some situations

we may want to conduct a separate analysis of leaf nodes and inner nodes. For example, we may wish to first determine which OTUs are detected while controlling FAR at some level q_1; then we may wish to conduct a second, separate analysis of taxa starting at the species level and continuing up the phylogenetic tree, while controlling the FAR of the second analysis at some level q_{-1}.

The procedures presented in Sections 2.2.4 and 2.2.5 do not guarantee that the FAR at each level l is controlled at level q_l because of the cumulative effect of D_{l-1} in (2.1), (2.4) and (2.5), which establishes a dependence between the nodes detected at each level. If we break this dependence by re-starting the counter at some level, it is then possible to separately control FAR above and below this level. Here, we illustrate our proposal by showing how to control FAR at level 1 to have value $\leq q_1$ while simultaneously controlling FAR at all remaining (higher) levels at value $\leq q_{-1}$. To accomplish this, we propose a two-stage procedure. At stage 1, we perform the step-down test for level 1 with thresholds $\{\alpha_{1,j}, j = 1, 2, \ldots, n_1\}$ that satisfy

$$\frac{\alpha_{1,j}}{1-\alpha_{1,j}} = \left(\frac{\sum_{k=1}^{j} \widetilde{\omega}_{1,(k)}}{\sum_{k=j}^{n_1} \widetilde{\omega}_{1,(k)}} \times q_1 \right) \bigwedge \frac{\tau_0}{1-\tau_0}.$$

Note that, if the same value of q_1 is used, these are the same thresholds for level 1 as the one-stage procedure described in the previous section. The use of weights $\{\widetilde{\omega}_{1,(k)}\}$ to account for multiplicity allows us to include in our list of detected nodes at level 1 all higher-level nodes that are detected after testing level 1 nodes. Thus, the FAR at level 1 is written as

$$\mathrm{FAR}_{\mathrm{otu}} = \mathbb{E}\left[\frac{\sum_{j=1}^{n_1} \omega_{1,j} V_{1,j}^m}{\left(\sum_{j=1}^{n_1} \omega_{1,j} R_{1,j} \right) \bigvee 1} \right],$$

At stage 2, we then apply the one-stage procedure proposed in Section 2.2.4 or 2.1.5 to the tree obtained by removing all leaves (OTUs) as well as those higher-level taxa that were detected at stage 1. In this tree, undetected nodes at level 2 are now the leaves, and the p-values for these new leaves are calculated by aggregating the p-values from the

undetected OTUs. We use the thresholds $\{\alpha_{l,j}, l = 2, \ldots, L, j = 1, \ldots, n_l^*\}$ that satisfy

$$\frac{\alpha_{l,j}}{1-\alpha_{l,j}} = \left(\frac{D_{l-1}^{\dagger} + \sum_{k=1}^{j} \widetilde{\omega}_{l,(k)}}{\sum_{k=j}^{n_l^*} \widetilde{\omega}_{l,(k)}} \times q_l \right) \bigwedge \frac{\tau_0}{1-\tau_0},$$

where $D_{l-1}^{\dagger} = \sum_{l'=2}^{l} d_{l'}^*$ differs from D_{l-1} in that $D_{l-1}^{\dagger}$ counts the detected nodes starting from the 2$^{\text{nd}}$ level. Using $D_{l-1}^{\dagger}$ in place of D_{l-1} cuts the dependence between level 1 and the remaining levels and also makes the thresholds more stringent if q_ls stay the same as those in the one-stage procedure. Thus, the FAR we wish to control at the remaining (higher) levels is

$$\text{FAR}_{\text{taxa}} = \mathbb{E}\left[\frac{\sum_{l=2}^{L} \sum_{j=1}^{n_l^*} \omega_{l,j} V_{l,j}^m}{\left(\sum_{l=2}^{L} \sum_{j=1}^{n_l^*} \omega_{l,j} R_{l,j} \right) \bigvee 1} \right],$$

where, as before, n_l^* excludes nodes detected using information from level 1 to $l-1$. Theorem 2.4 states that this two-stage procedure serves our purpose.

Theorem 2.4. Under Conditions (C2) and (C3) in Theorem 2.1, the above two-stage procedure ensures that $\text{FAR}_{\text{otu}} \leq q_1$ and $\text{FAR}_{\text{taxa}} \leq q_{-1} = \sum_{l=2}^{L} q_l$.

The proof of Theorem 2.4 is proved in Appendix A.5. Note that the choice of $(q_1, q_2, \ldots, q_L)$ is at the user's discretion and not necessary to match those in the one-stage procedure. For example, we can set $q_1 = q_{-1} = 5\%$ and choose $q_l = q_{-1} n_l / \left(\sum_{l'=2}^{L} n_l \right)$ for $l = 2, \ldots, L$.

Although we have presented the example in which the FAR for leaf nodes are controlled separately from the FAR for inner nodes, in principal the approach described here could be used to divide the nodes into two groups at any level, simply by choosing where to zero the counter in D_l. We could even apply the approach recursively to control FAR for more than two sets of levels, if desired.

2.3 Simulation Studies

We conducted simulation studies to assess the performance of our bottom-up tests, and to compare with three competing approaches: (1) the naïve approach that calculates the p-value for an inner node by aggregating p-values from *all* leaf nodes that are its descendents, using Stouffer's Z-score method and applies the BH procedure on the collection

of p-values from all nodes; (2) the top-down approach of Yekutieli (2008) as implemented in the R package `structSSI` (Sankaran and Holmes, 2014), with p-values for inner nodes calculated in the same way as in the naïve approach; and (3) the conjunction-null test that assigns a p-value to an inner node by the *largest* p-value from all offspring nodes (equivalently, the largest p-value from all corresponding leaf nodes) and applies the BH procedure as in the naïve approach. All methods take p-values at leaf nodes as input. The nominal level for all error rates was set to 10%.

All simulations were conducted under the conjunction null. We first selected a number of inner or leaf nodes to be driver nodes. Under the conjunction null, all offspring of driver nodes (including all its leaf nodes) are associated with the trait of interest. We independently sampled p-values for associated leaf nodes from distributions that have enriched probability at values close to zero. We used the Beta distribution Beta$(1/\beta, 1)$ where $\beta > 1$, which has a relatively heavy right tail (Figure A.1) and mimics the empirical distributions of p-values observed in the IBD data (Figure A.2). To assess the robustness of our results, we also considered sampling p-values from a Gaussian-tailed model frequently used to study the performance of FDR procedures (Storey, 2002; Barber and Ramdas, 2017; Javanmard and Montanari, 2018), which has a smaller right tail (Figure A.1). Specifically, we first drew values $X_{l,j} \sim N(\beta, 1)$ and then obtained the p-value $p_{l,j} = 1 - \Phi(X_{l,j})$, where Φ is the standard normal cumulative distribution function. In both models, β characterizes the effect size of the trait on the microbiome. For all simulations we assumed that all leaf nodes that were not descendants of driver nodes were null with p-values sampled independently from the U$[0, 1]$ distribution.

We considered three tree structures, shown in Figure 2.2. The first is a complete binary tree with 2 children for each inner node and 10 levels, which has 1023 nodes of which 512 are leaves. The second is a complete 'bushy' tree with 10 children for each inner node and 4 levels, which has 1111 nodes of which 1000 are leaves. The third is a real phylogenetic tree (Halfvarson et al., 2017) that we apply our methods to in Section 2.4. This tree has 8 levels, 249 inner nodes, and 2360 leaf nodes, with large variation in the number of child nodes at different inner nodes. It is incomplete, having extensive ($>$ 50%) missing assignments at the genus and species levels, and a few at the family level.

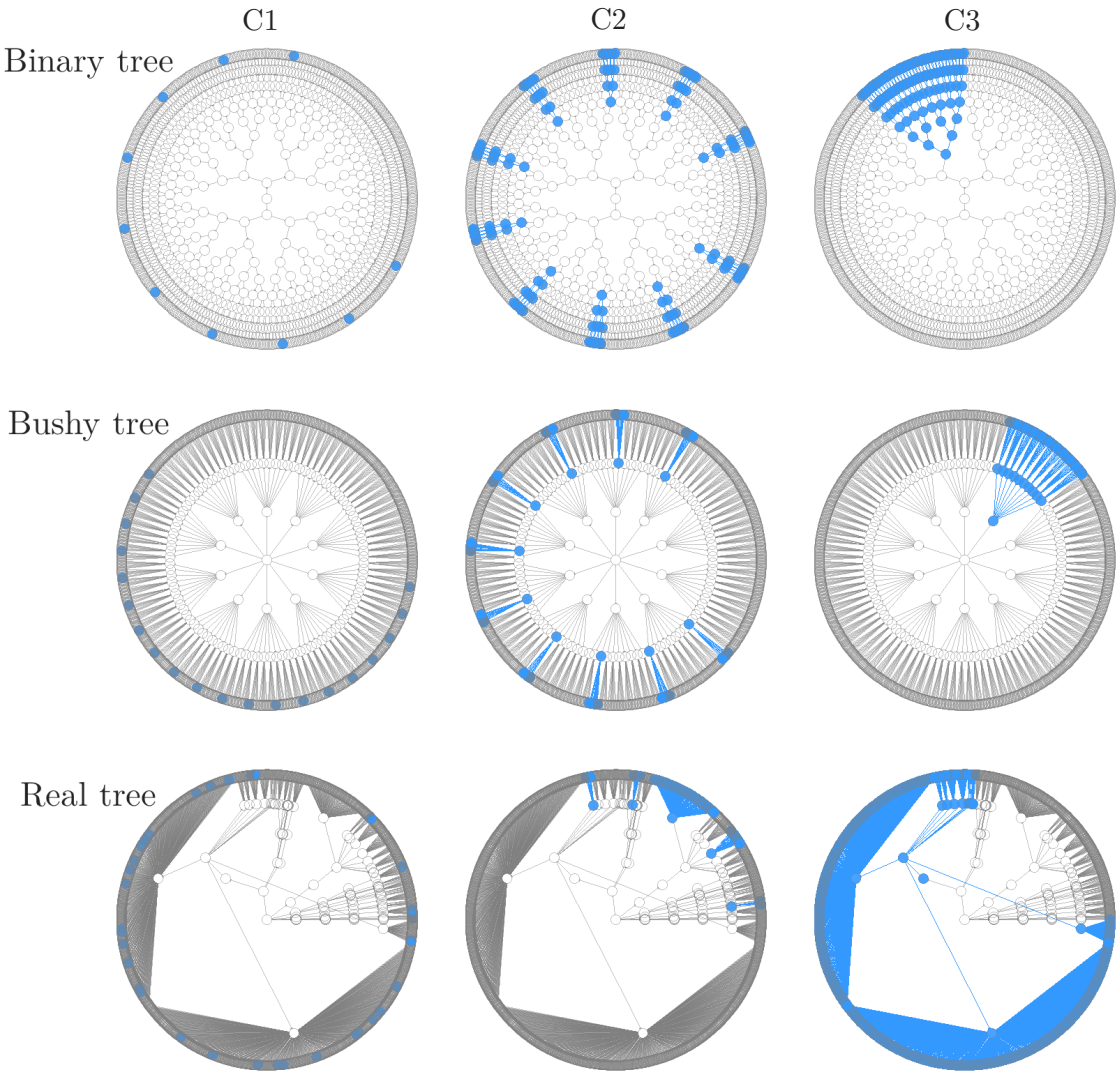

Figure 2.2: The three tree structures (binary, bushy, and real) and three causal patterns (C1, C2, and C3) for simulation studies. The real phylogenetic tree structure was obtained from the IBD data and, for simplicity of exposition, skipped the genus and species levels which have extensive missing assignments. The root node is always located at the center of each tree and the leaf nodes are represented by the outermost ring. The top of each blue subtree is a designated driver node, which can be an inner or leaf node.

For each tree structure, we then considered three causal patterns, differentiated by the level of the selected driver nodes (Figure 2.2). The first pattern (C1) is characterized by sparse driver nodes located at the leaf nodes; the second (C2) by several driver nodes located at an intermediate level, chosen so that ∼10% of leaf nodes were associated; and the third (C3) by a single driver node at a higher level, inducing association with a large subtree. In particular, for C1 we randomly selected 10/20/36 leaf nodes for the binary/bushy/real trees. For C2, we randomly chosen 10 out of 64 nodes at level 4 for the binary tree, 10 out of 100 nodes at level 2 for the bushy tree, and 5 out of 48 nodes at the family level (level 4) for the real tree. For C3, we randomly picked one node at level 7 for the binary tree, one node at level 3 for the bushy tree, and for the real tree, the class *Clostridia* from level 6 (covering ∼80% of all leaf nodes). Each of the $3 \times 3 = 9$ scenarios was replicated 1000 times.

2.3.1 Error Rates

We evaluated each procedure by calculating its FAR (under the modified null), FDR (under the global null), and its FDRc (under the conjunction null). The FAR of the top-down and naïve approaches were calculated by using the list of detected nodes (as determined by Benjamini and Hochberg (1995) for the naïve method and Yekutieli (2008) for the top-down method), then proceeding level by level as if these detections were the result of tests of the modified null hypothesis. The FAR was then calculated using the weighted procedures of Section 2.2.4 for the binary and bushy trees and Section 2.2.5 for the real tree.

Figure 2.3 displays these results for the nine ($= 3 \times 3$) scenarios we considered, for the simulations that used the beta distribution for non-null leaves. In all 9 scenarios our methods, whether unweighted or weighted, always controlled FAR. In contrast, the naïve and top-down methods typically had inflated FAR, with the most severe inflation occurring in C1, where the driver nodes were simulated exclusively at leaf nodes, causing many higher-level nodes to be falsely detected by these methods. As expected, our methods also controlled FDR. The top-down method, although designed to control FDR,

still yielded slightly inflated FDR occasionally; this is likely due to violation of the independence assumption between the p-value at a node and the p-values of all its ancestors, which is required by the top-down method (Yekutieli, 2008). The naïve method always controlled FDR because the BH procedure is known to be robust to such positive correlations. Despite a lack of theoretical results, we found that our methods (especially the weighted approach of Section 2.2.4) controlled FDRc reasonably well in all scenarios we considered. The naïve and top-down methods typically had inflated FDRc, and FDRc for these methods resembled their FAR, consistent with the notion that FAR approximates FDRc. The conjunction-null test controlled all error rates, as expected. Figure A.3 shows the same patterns of FAR, FDR, and FDRc for simulations based on the Gaussian-tailed model.

2.3.2 Accuracy and Pinpointing Driver Nodes

As our simulations were conducted under the conjunction null hypothesis, the driver nodes and all of its descendants are the truly associated nodes. We measured accuracy by calculating a Jaccard similarity between the set of truly associated nodes and the nodes that are detected, for each method. We used a *weighted* Jaccard similarity to account for the branching-tree topology of the hypotheses we test, because detecting a node with offspring implies the offspring are detected in some sense, even if they were not individually detected. For example, identifying a genus as being associated with the trait of interest implies the species and OTUs that belong to this genus are associated, even if we did not detect them (and hence are not included in the list of detected nodes). For this reason, we calculated the Jaccard similarity by weighting each node by the number of leaf nodes that are its descendants. The weighted Jaccard similarity is then the sum of the weights of correctly-detected nodes, divided by the total weight assigned to either detected or truly associated nodes.

Examining Figure 2.4 we see that the weighted bottom-up approach has the best or second-best accuracy (as measured by the weighted Jaccard similarity) in all cases we examined, making it the best overall choice. The test of the conjunction hypothesis slightly

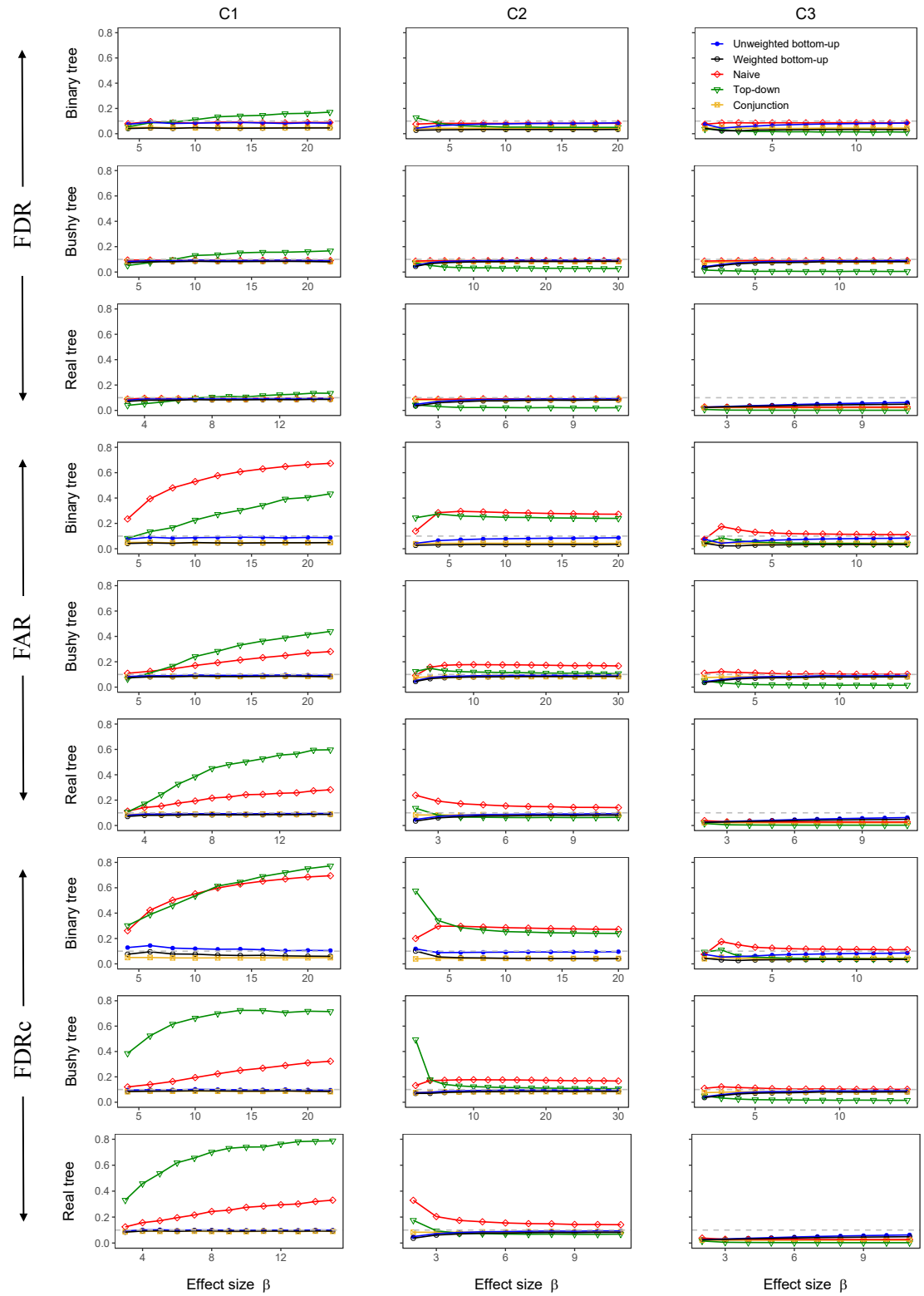

Figure 2.3: Error rates for testing all nodes in the tree. The non-null p-values at leaf nodes were simulated from the beta distribution.

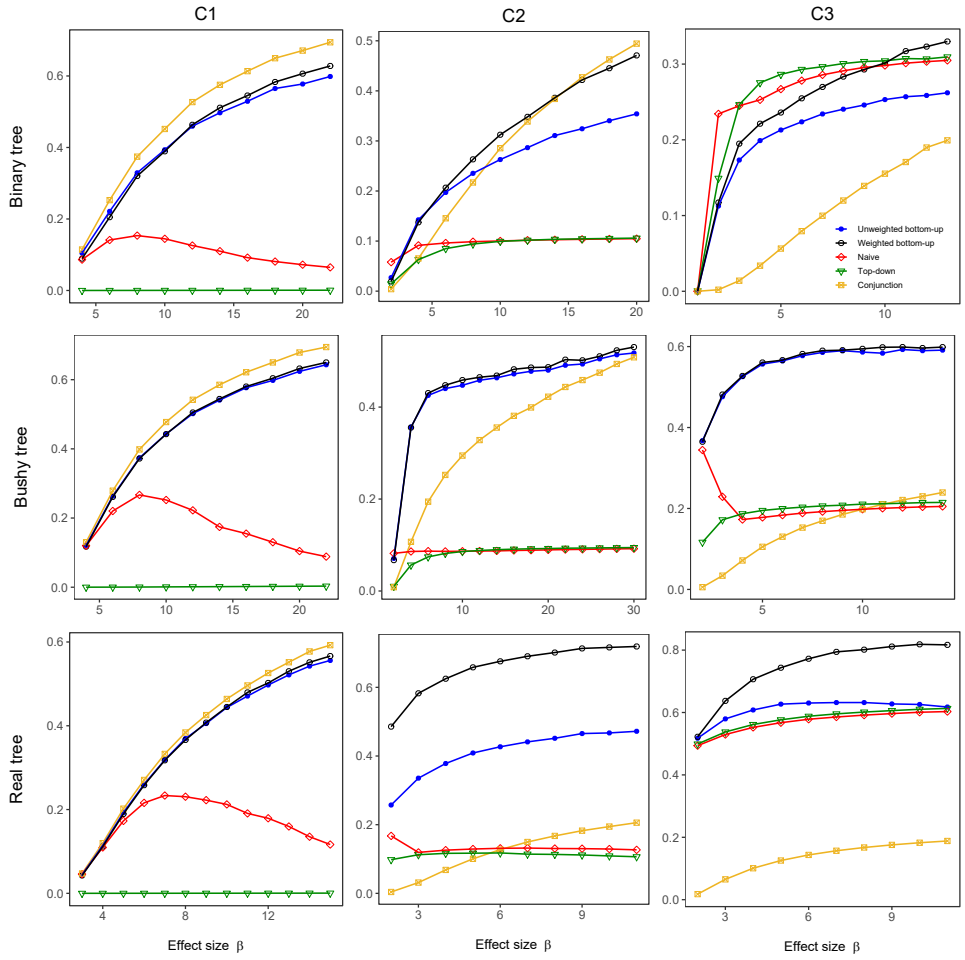

Figure 2.4: Accuracy (weighted Jaccard similarity) for detecting all associated nodes (including the driver nodes and all of their descendants at all levels). The non-null p-values at leaf nodes were simulated from the beta distribution.

outperforms our bottom-up tests when only leaf nodes (OTUs) are truly associated (simulation C1) as its natural conservatism is an advantage in this case. As soon as inner nodes are truly associated, as in C2 or C3, the conjunction test becomes very conservative. The bottom-up approaches performed best for C2 and C3 except for the binary tree in C3, where the naïve and top-down approaches performed better at low parameter values. When considering these results, it should also be noted the naïve method (but not the top-down method) had elevated FAR and FDRc for this simulation. In general, the performance of our weighted bottom-up procedure was either equivalent to or superior to the unweighted procedure, presumably reflecting the ability of the weighted procedure to include all nodes that are identified when all offspring of some node are detected.

Our methods are most different from existing methods in their ability to pinpoint driver nodes. We say a driver node is "pinpointed" if it is detected to be associated *and*

none of its ancestors are detected. We evaluated the percentage of driver nodes that were pinpointed and showed the results in Figures 2.5 and A.5. In general, our weighted and unweighted methods pinpointed a similar number of driver nodes, and both detected many more driver nodes than the naïve, top-down, and conjunction-null methods. The naïve method pinpointed some driver nodes when the association signals were weak, but inevitably detected their ancestors as the signals became stronger. By definition, the top-down method must fail to pinpoint any driver node since it only tests nodes below the root node if the root node is detected. Note that the percentage of driver nodes pinpointed by our methods sometimes decreased as the effect size increased, because more (but not all) descendants were detected and removed from the statistic for the driver nodes, which thus aggregated less information. For the beta-based simulations, undetected driver nodes remained a possibility regardless of the effect size, as the beta distribution always generates a non-negligible portion of p-values that were close to one even when the effect size was extremely large. For the Gaussian-tailed simulation, all driver nodes were eventually detected, as large p-values became more infrequent as the effect size increased. We note that the conjunction-null test can easily fail to detect higher-level nodes (including driver nodes) when some offspring of these nodes have large p-values, as in our beta-based simulations. Additionally, we note that the good accuracy (as measured by the weighted Jaccard similarity) of our approaches is related to their ability to pinpoint driver nodes.

2.4 IBD Data

IBD is a chronic disease accompanied by inflammation in the human gut. The two most common subtypes are ulcerative colitis (UC) and Crohn's disease. Halfvarson et al. (2017) investigated the longitudinal dynamics of the microbial community in an IBD cohort of 60 subjects with UC and 9 healthy controls. The microbial community was profiled by sequencing the V4 region of the 16S rRNA gene. Sequence data were processed into an OTU table through the QIIME pipeline. Our goal was to identify taxa that have differential abundance between the UC and control groups at baseline.

We removed OTUs that were present in fewer than 10 samples and dropped 4 OTUs

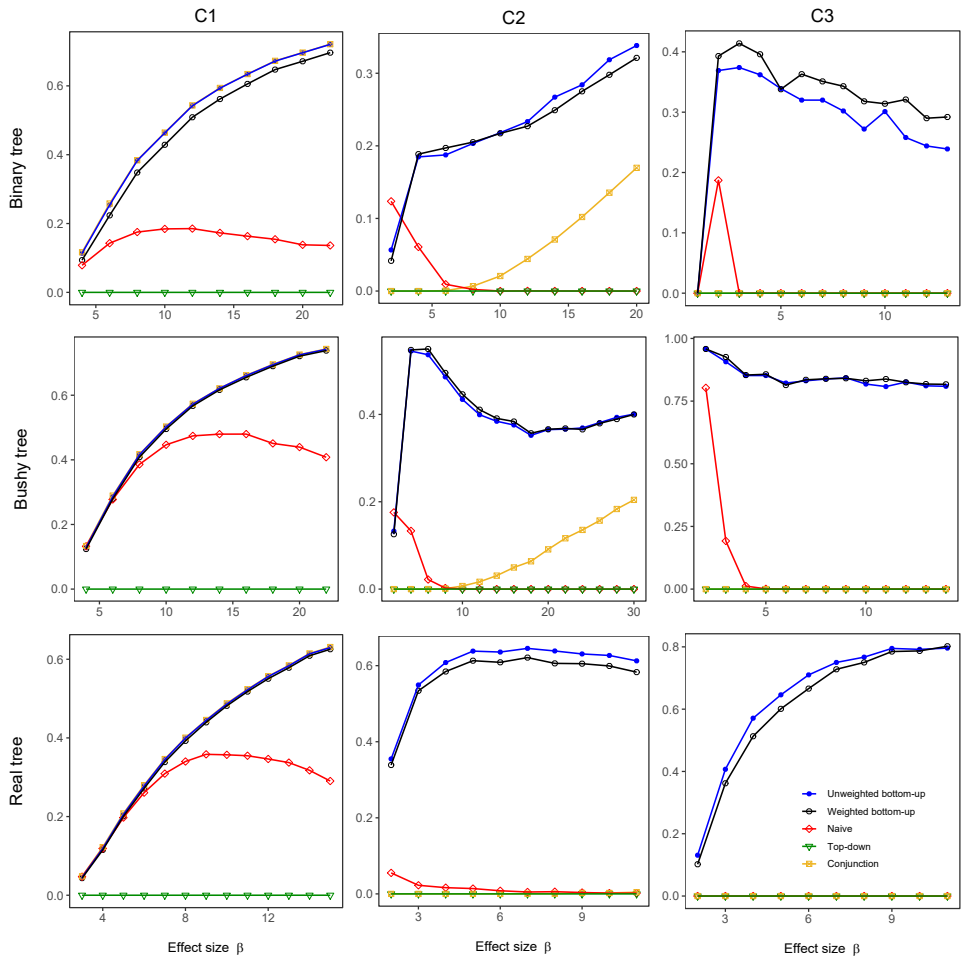

Figure 2.5: Percentage of driver nodes that were pinpointed. The non-null p-values at leaf nodes were simulated from the beta distribution.

that failed to be assigned any taxonomy. The assigned taxonomy grouped the 2360 OTUs into 249 taxonomic categories (i.e., inner nodes) corresponding to kingdom, phylum, class, order, family, genus, and species levels. Note that 15.2%, 56.9%, and 91.3% OTUs have missing assignment at the family, genus, and species level, respectively. As there were no obvious confounders provided with these data, we used the Wilcoxon rank-sum test to compare the OTU frequencies between case and control groups to obtain p-values for each OTU.

We applied the two bottom-up methods with the nominal FAR of 10% as well as the naïve and top-down methods with the nominal FDR of 10%. The detected taxa can be visualized in Figure 2.6. The weighted bottom-up test identified 127 OTUs, 6 species, 9 genera, 7 families, 5 orders, 6 classes and 1 phylum, among which the driver taxa were pinpointed at phylum *Verrucomicrobia*, classes *Chloroplast*, *Clostridia*, *Coriobacteriia*, *Erysipelotrichi* and *RF3*; families *Prevotellaceae* and *S24-7*; genera *Morganella* and [*Prevotella*]; and species *ovatus* and *radicincitans*; see Table A.1 for more details. The unweighted procedure yielded a very similar list of driver taxa. In contrast, both the naïve and top-down methods identified the root node and many taxa at high levels, suggesting their inability to pinpoint the real driver taxa. In addition, the top-down method did not detect some lower-level taxa in phylum *Proteobacteria* that were detected by all other methods. The conjunction test detected 142 OTUs but only 5 taxa, each of which contained a single OTU. All these results are consistent with the findings of our simulation studies.

We also applied the two-stage (weighted) bottom-up procedure to control FAR separately at the OTU level and the remaining taxa levels. We considered splitting the overall FAR 10% into 5% and 5% for OTU and taxon analyses. At stage 1, we detected 79 OTUs and also included in our detection list 1 species, 1 genus, 1 order, and 1 class because they all contain only 1 OTU and those OTUs were detected; among these OTUs and taxa we can declare we control FAR at 5% ($\sim$ 4 OTUs or taxa). At stage 2, we detected 4 species, 4 genera, 5 families, 4 orders, 3 classes, and 2 phyla, among which we can declare we control FAR at 5% ($\sim$ 1 taxon). The detected driver taxa were phyla *Bacteroidetes* and *Verrucomicrobia*, classes *Chloroplast*, *Coriobacteriia* and *Erysipelotrichi*,

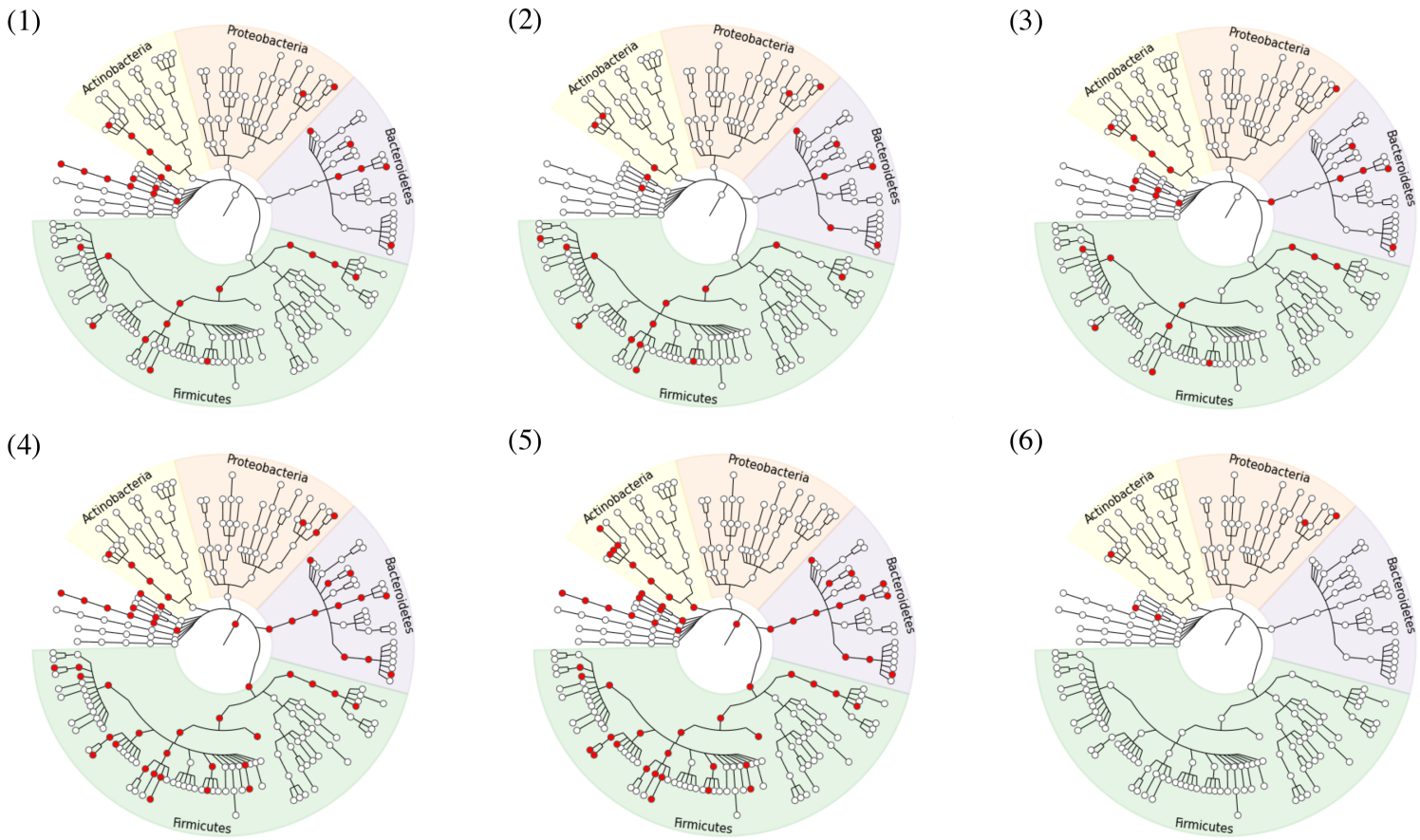

Figure 2.6: Taxa (marked in red) detected to be differentially abundant between the UC and control groups in the IBD data by (1) the (one-stage) weighted bottom-up method, (2) the unweighted bottom-up method, (3) the two-stage, weighted bottom-up procedure with nominal FAR 5% and 5% for OTUs and taxa, respectively, (4) the naïve method, (5) the top-down method, and (6) the conjunction-null test. The levels from the center outward are kingdom, phylum, class, order, family, genus and species. The OTU level is supposed to be located at the outermost layer and has been omitted to simplify the figure. The plots were generated using GraPhlAn (`http://huttenhower.sph.harvard.edu/graphlan`).

order *Clostridiales*, and species *radicincitans*. Note that the two FARs (for OTUs and taxa) do not have to add up to 10% nor be equal.

2.5 Discussion

In this chapter, we presented a bottom-up approach to testing hypotheses that have a branching tree dependence structure. These procedures test hypotheses in a tree level by level, starting from the bottom and moving up, rather than starting from the top and moving down. We developed a novel modified null hypothesis, which is more suitable for our goal of detecting nodes in which a *dense* set of child nodes are associated with the trait of interest. Accordingly, we developed a novel error criterion, the FAR, and provided procedures that we proved control FAR. Our simulation studies confirmed the control of FAR and demonstrated good performance of our methods compared to existing methods using a measure of accuracy based on a weighted Jaccard similarity. Further, our bottom-up methods are more successful at pinpointing driver nodes, offering highly interpretable results, while the existing methods frequently fail at this task. Finally, although our methods were not designed to control FDRc, our simulations showed that use of Stouffer's Z score to combine information leads to approximate control of FDRc as well.

Our methods can easily be extended to very general tree structures. We can easily handle trees in which the leaf nodes are not all at level 1. With some modifications, our methods can also be applied to trees with multiple (correlated) root nodes such as trees generated by pathways, by using our bottom-up testing procedure up to the level right below the root level and applying to the root level the standard BH procedure, which is robust to positive correlations. We expect this modified procedure to control FAR (and hence FDR and, approximately, FDRc).

Although our approach is very general, it does have some limitations or aspects that could benefit from further development. First, we treated p-values at leaf nodes with equal weight. In some applications, different leaf nodes may have varying importance and may be weighted differently. Second, we partitioned the total error rate q into

$q_1, \ldots, q_L$ in proportion to the number of nodes at each level. It is unclear what the optimal partition would be. We have done a sensitivity analysis comparing our partition with a number of alternative ones and found that our partition achieved the most robust performance in detecting associated nodes as well as driver nodes over all scenarios that we have considered in simulation (Figure A.6). It is also of interest to consider alternative partitioning that can improve performance at pre-specified levels of particular importance if, for example, finding genera that were associated with a trait was of particular interest. Further, we assumed independence between null leaf nodes because it is required by both Stouffer's method for combining p-values and the step-down procedure for controlling the error rate of decisions. It may also be of interest to extend our methods to account for correlations between leaf nodes, e.g., correlations between (null) OTUs.

Our methods have been implemented in the R package `BOUTH` (BOttom-Up Tree Hypothesis testing), available on GitHub links.. Our program is computationally efficient as it only involves building the tree structure, calculating the thresholds, aggregating p-values for parent nodes, and sorting p-values. For example, for the one-stage weighted procedure on the IBD data, it took 1.3 seconds on a laptop with a 2.5 GHz Intel Core i7 processor and 8 GB RAM.

Chapter 3

Controlling Type-I Monte Carlo Error Rate in Resampling-Based Multiple Testing

3.1 Introduction

In this chapter, we discuss the problem of controlling family-wise type-I Monte Carlo error rate (MCER-I) in resampling-based multiple testing. In Section 3.2.1, we discuss a potentially useful approach motivated by bootstrap heuristics, as we observe that a decision on ideal p-values usually matches the majority voting under bootstrap empirical distribution. However, we notice that the error rate cannot be controlled in some bad scenarios. In Section 3.2.2, we propose a two-step approach that divides the problem into two sub-problems that are more tractable. This two-step approach achieves better error rate control and is more computational efficient. Hence we recommend this method. In Section 3.3, properties of the proposed methods are comprehensively studied through numerical experiments. Performances of our proposed methods are demonstrated through a prostrate cancer gene expression dataset in Section 3.4.

3.2 Methods

We denote the m (e.g., $m = 1000$) null hypotheses by $H_{1,0}, H_{2,0}, \ldots, H_{m,0}$, and the *ideal* p-values for testing these hypotheses by $p_1^*, p_2^*, \ldots, p_m^*$. Consider that the BH procedure with nominal FDR q is applied to the ideal p-values. The ideal p-values are sorted in an ascending order $p_{(1)}^* \leq p_{(2)}^* \leq \cdots \leq p_{(m)}^*$. If k is the largest integer such that $p_{(k)}^* \leq q \times k/m$, then the null hypotheses corresponding to $p_{(1)}^*, \ldots, p_{(k)}^*$ are rejected and the remaining hypotheses are accepted. Thus, we call $\tau^* = q \times k/m$ that separates rejected and accepted p-values the BH *cutoff*. If τ^* is known *a priori*, we can rewrite the original multiple testing problem to be

$$H_{i,0} : p_i^* > \tau^* \qquad \text{versus} \qquad H_{i,1} : p_i^* \leq \tau^*.$$

In reality, τ^* is unknown. In space $[0,1]^m$, both events $\{p_i^* > \tau^*\}$ and $\{p_i^* \leq \tau^*\}$ have quite complex geometric structures. Testing these hypotheses would be a major challenge in our methods development.

Before we give our solutions to this problem, we first describe the input of the whole

procedure. Suppose that n permutation replicates have been collected. For every hypothesis $H_{i,0}$ and every permutation replicate indexed by j, we can obtain the test statistic T_{ij} and then the corresponding exceedance indicator $I_{ij} = \mathbb{I}(|T_{ij}| \geq |t_i|)$, where t_i is the observed statistic. Our input is the matrix $\boldsymbol{I} = \{I_{ij}\}_{m\times n}$. From this matrix, we can obtain the total number of exceedances for each test, $X_i = \sum_{j=1}^{n} I_{ij}$. As the replicates are sampled independently from the null distribution, we have $I_{ij} \sim$ Bernoulli(p_i^*) and then $X_i \sim$ Binomial(n, p_i^*).

3.2.1 The Empirical Strength Probability (ESP) Approach

The naïve ESP procedure

In this section, we first consider an approach that directly uses the bootstrap empirical distribution of permutation p-values. We rewrite the hypothesis testing problem $H_{i,0} : p_i^* > \tau^*$ versus $H_{i,1} : p_i^* \leq \tau^*$ to be:

$$H_{i,0} : \mathbf{p}^* \in \mathcal{A}_i \qquad \text{versus} \quad H_{i,1} : \mathbf{p}^* \in \mathcal{R}_i,$$

where $\mathcal{A}_i$ and $\mathcal{R}_i$ represent the acceptance region and rejection region respectively for classifying p_i^*, with standard BH procedure under FDR nominal level q. The acceptance region $\mathcal{A}_i = [0,1]^m \setminus \mathcal{R}_i$. In general, for two distinct $i_1 < i_2$, the corresponding rejection regions $\mathcal{R}_{i_1} \neq \mathcal{R}_{i_2}$, but their structures are highly similar. There exists an one-to-one mapping f such that $\mathcal{R}_{i_1} = f(\mathcal{R}_{i_2})$: basically any $\mathbf{p}^* = (p_1^*, \cdots, p_{i_1}^*, \cdots, p_{i_2}^*, \cdots p_m^*) \in \mathcal{R}_{i_1}$ becomes $f(\mathbf{p}^*) = (p_1^*, \cdots, p_{i_2}^*, \cdots, p_{i_1}^*, \cdots p_m^*) \in \mathcal{R}_{i_2}$ by switching the two coordinates. An example of rejection and acceptance regions when $m = 2$ is illustrated in Figure 3.1 (a).

Let B be the number of bootstrap samples for this ESP approach. For each $b = 1, 2, \ldots, B$, we propose to obtain the b-th bootstrap sample of $(X_1, \ldots, X_m)$, denoted as $(X_1^{(b)}, X_2^{(b)}, \ldots, X_m^{(b)})$, by sampling n columns of $\boldsymbol{I}$ with replacement to obtain a new matrix $\boldsymbol{I}^{(b)}$ and to sum up the rows. To improve computational efficiency, we adopted a sampling strategy based on column blocks (so the columns in one block are in or out together). Using the r-column blocks can reduce the time of bootstrap sampling by r fold,

which is very useful when n and B are very large. After sampling $(X_1^{(b)}, X_2^{(b)}, \ldots, X_m^{(b)})$, we construct the bootstrap p-values from each bootstrap sample:

$$\widehat{p}_i^{(b)} = \frac{X_i^{(b)} + 1}{n + 1},$$

where $i = 1, \ldots, m$. Then we apply the standard BH procedure to $(\widehat{p}_1^{(b)}, \widehat{p}_2^{(b)}, \ldots, \widehat{p}_m^{(b)})$ and obtain the decisions $(\widehat{D}_1^{(b)}, \widehat{D}_2^{(b)}, \ldots, \widehat{D}_m^{(b)})$. We assess whether each of $(H_{1,0}, H_{2,0}, \ldots, H_{m,0})$ can be stably rejected over the majority of B bootstrap samples.

More specifically, define ω_i to be the proportion of B bootstrap samples in which $H_{i,0}$ is accepted:

$$\omega_i = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left(\widehat{D}_i^{(b)} = acceptance\right).$$

A small ω_i value shows strong evidence for rejecting $H_{i,0}$. Such an ω_i is called empirical strength probability (ESP) in literature (Liu and Singh, 1997; Davison et al., 2003) because the strength in support of the null hypothesis comes from the observed data, which is known to behave like p-value asymptotically (Liu and Singh, 1997) under mild conditions. In particular, for $H_{i,0}$ which should not be rejected based on the ideal p-value, ω_i is expected to asymptotically follow uniform distribution U$[0, 1]$, or be more conservative than U$[0, 1]$ (i.e., $\limsup_{n\to\infty} \Pr(\omega_i \leq t) \leq t$). Therefore, we propose to calculate $\omega_1, \omega_2, \ldots, \omega_m$ and then performs Holm's procedure (Holm, 1979) on them to asymptotically control the FWER by α.

There are two major advantages of this algorithm. First, it enjoys significantly improved power (i.e., number of discoveries) compared to the existing methods. Second, the ESP approach provides a confidence measure for each rejection it makes. A smaller ω_i indicates a higher confidence in rejecting the i-th hypothesis.

However, the ESP approach may occasionally inflate the family-wise MCER-I in finite samples. In some bad scenarios, it can reject more hypotheses than it should be. It is known that the ESP value ω_i can be substantially biased (Efron and Tibshirani, 1998; Davison et al., 2003) when the null region $\mathcal{A}_i$ has nonsmooth boundary points and the ideal p-values are close to at least one of these nonsmooth corners. Particularly in our case, when some ideal p-values are close to multiple BH check points (i.e., $q/m, 2q/m, \ldots, q$), it

will require an extremely large amount of permutations to ensure that this approximation is sufficiently accurate. It becomes even worse if more ideal p-values are near the BH check points. The BH decision boundary for testing p_i^* is a surface $\mathcal{S}_i = \mathcal{R}_i \cap \overline{\mathcal{A}_i}$. The $\overline{\mathcal{A}_i}$ means the closure of $\mathcal{A}_i$. Every point on the surface $\mathcal{S}_i$ can be classified into two categories: $\mathcal{S}_{i,0}$ is the collection of ordinary (boundary) points such that $\mathcal{S}_i$ is differentiable at the point; and $\mathcal{S}_{i,1}$ is the collection of singular (boundary) points such that the boundary surface cannot take derivatives there. In Figure 3.1, there are only 2 singular points which are the two corners. When $m \geq 3$, the number of singular points would be infinity.

Next we consider classification of singular points. However, this question would become very complicated under the general dependency structure among m tests. Here we assume that all tests are independent. We define a metric V for any $\mathbf{p} \in$ the collection of boundary points:

$$V(\mathbf{p}) = \lim_{r \to 0^+} \frac{\text{vol}(B(\mathbf{p}, r) \cap \mathcal{R}_i)}{\text{vol}(B(\mathbf{p}, r) \cap [0,1]^m)},$$

where $B(\mathbf{p}, r)$ means an m-dimensional ball centering at $\mathbf{p}$ with radius $= r$, and vol means the volume of an closed object. Under this independence assumption and in the asymptotic sense, the $V(\mathbf{p})$ is proportional to the probability of i-th test falling into the rejection region given ideal p-values $= \mathbf{p}$. When $\mathbf{p}$ is an ordinary boundary point, $V(\mathbf{p})$ is always $1/2$. Then we use the metric V to further classify singular points into two classes. The type-I singular points satisfy $V(\mathbf{p}) > 1/2$. It is easy to verify that $\text{P}s_1$ in Figure 3.1 (a) is such a singular point. When the ideal p-values are in $\mathcal{A}_i$ but too close to a type-I singular point, typically $\Pr(\omega_i \leq t) > t$ and the type-I error of testing $H_{i,0} : p_i^* \in \mathcal{A}_i$ is inflated. The type-II singular points satisfy $V(\mathbf{p}) \leq 1/2$. The other singular point $\text{P}s_2$ in Figure 3.1 (a) belongs to this class. We should not observe type-I error inflation, if the ideal p-values we mentioned above is only close to type-II singular points.

The other limitation of this algorithm is related to its computational complexity. Because the Holm's procedure requires us to compare ESP values to the thresholds $\{\alpha/m, \alpha/(m-1), \cdots, \alpha\}$. However, ESPs can only be chosen from discrete values $0, 1/B, 2/B, \cdots, 1$. Therefore B should be at least in the order of $O(m)$ to ensure that the estimates of ESP values are accurate enough.

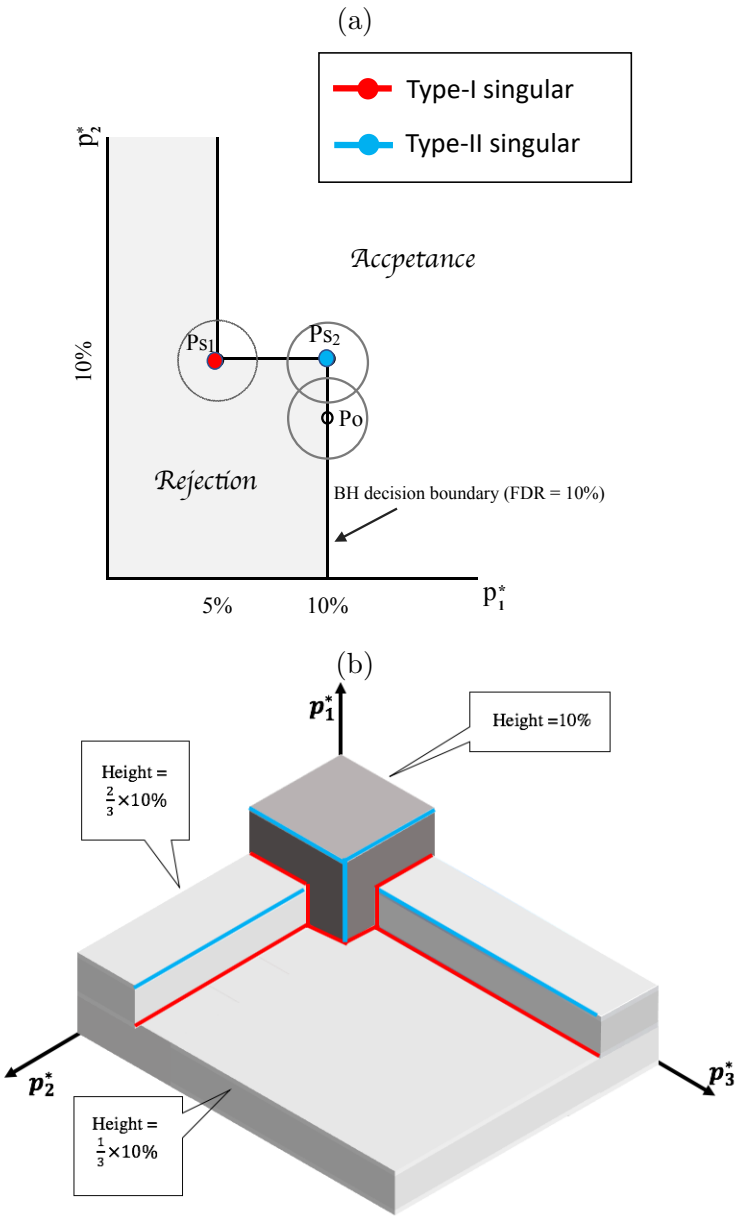

Figure 3.1: An illustrative example of ESPs and regions classified by BH decisions. In figure (a) we consider the decision for p_1^* when there are two tests. According to Liu and Singh (1997), the ESP value ω_1 would asymptotically follow U$[0, 1]$ distribution if ideal p-values are chosen from ordinary boundary points, for instance, "Po". At these points the BH decision boundary is smooth. In contrast, the decision boundary is not smooth at the type-I singular point"Ps_1". The ω_1 based on a limited number of permutations would be anti-conservative to U$[0, 1]$ if the ideal p-values are close to "Ps_1". Under that scenario, bootstrap samples are more likely to fall in the rejection region, which causes the ESP value to be enriched around 0. The other singular point in figure (a) is "Ps_2". However, it is a type-II singular point and hence will not cause inflated ω_1. In figure (b) we consider the decision for p_1^* when there are three tests. The combination of gray cuboids represents the rejection region by BH for p_1^*. All type-I and type-II singular points are colored by red and blue respectively.

Algorithm 2 The Empirical Strength Probability (ESP) approach

The naïve ESP procedure

1. For each bootstrap sample $(X_1^{(b)}, X_2^{(b)}, \ldots, X_m^{(b)})$, calculate bootstrap p-values $(\widehat{p}_1^{(b)}, \widehat{p}_2^{(b)}, \ldots, \widehat{p}_m^{(b)})$, and apply the BH procedure to obtain decisions $(\widehat{D}_1^{(b)}, \widehat{D}_2^{(b)}, \ldots, \widehat{D}_m^{(b)})$
2. Calculate $(\omega_1, \omega_2, \ldots, \omega_m)$, treat them as p-values, and apply the Holm's procedure: for ordered $\omega_{(1)} \leq \omega_{(2)} \leq \cdots \leq \omega_{(m)}$, find k to be the largest integer such that $\omega_{(i)} \leq \alpha/(m-i+1)$ for all $i \leq k$, and reject the corresponding k hypotheses $H_{(1),0}, \ldots, H_{(k),0}$

The validation test for bad scenarios
Assume m^+ is the number of rejected hypotheses with ESP procedure. The hypotheses with respect to $X_{(1)}, X_{(2)}, \ldots, X_{(m^+)}$ are rejected.

1. Begin with $l = 1$.
2. Stop testing if $l > m^+$. For $l = 1, 2, \ldots, m^+$, calculate $\widehat{q}_{(l)}$, the l-th smallest q-value of the observed permutation p-value $\widehat{\boldsymbol{p}}$, using formula (3.1).
3. Sample the empirical version of permutation p-values, $\widehat{\boldsymbol{p}}^{(b)}$, which is defined in ESP procedure.
4. Find $\ddot{\boldsymbol{p}} = (\ddot{p}_1, \ddot{p}_2, \ldots, \ddot{p}_m) = (\max(\widehat{p}_1, q \times 1/m), \max(\widehat{p}_2, q \times 2/m), \ldots, \max(\widehat{p}_m, q))$ which is the candidate singular point near $\widehat{\boldsymbol{p}}$. Create the empirical distribution of permutation p-values centering at $\ddot{\boldsymbol{p}}$ by the collection of $\ddot{\boldsymbol{p}}^{(b)} = (\widehat{p}_1^{(b)}, \widehat{p}_2^{(b)}, \ldots, \widehat{p}_m^{(b)}) - (r_1, r_2, \ldots, r_m) + (\ddot{p}_1 - \widehat{p}_1, \ddot{p}_2 - \widehat{p}_2, \ldots, \ddot{p}_m - \widehat{p}_m), b = 1, 2, \ldots, B$, where $r_1, r_2, \ldots, r_m$ are randomly sampled from U$[0, \frac{1}{n+1}]$ distribution for smoothing.
5. Find the empirical distribution of $\ddot{q}_{(l)}^{(b)}$, the l-th smallest q-value of each $\ddot{\boldsymbol{p}}^{(b)}$ using formula 3.1. Calculate the validation test p-value $p_l^{\text{val}} = \#\{b \in \{1, 2, \ldots, B\}, \text{ such that } \ddot{q}_{(l)}^{(b)} \leq \widehat{q}_{(l)}\}/B$.
6. If $p_l^{val} \leq \alpha$, reject the test that corresponds to $X_{(l)}$ and continue to step 2 with the updated $l = l+1$. Otherwise stop any further testing and claim tests that correspond to $X_{(l)}, X_{(l+1)}, \ldots, X_{(m)}$ are accepted.

Validation test for ESP procedure

Our next question is that can we fix the inflated family-wise MCER-I with the naïve ESP procedure? A possible solution is to consider corrected ESP values. There have been discussions on performing corrections on ESP values when the boundary of null region being tested is non-smooth. Liu and Singh (1997) suggested using $\max\{\text{ESP}, p_1, p_2, \cdots, p_s\}$ as the corrected ESP value, where there are finite singular points $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_s$ on the boundary of null region and for $i \in \{1, 2, \cdots, s\}$, p_i is the p-value of testing point-wise nulls $H_0 : \mathbf{p}^* = \mathbf{a}_i$ versus $H_1 : \mathbf{p}^* \neq \mathbf{a}_i$. at those singular points. However, in our case when $m \geq 3$, the number of both type-I and type-II singular points can reach infinity. See an example for $m = 3$ in Figure 3.1, where all points from those colored edges are singular. It is impossible to enumerate all p-values from testing those singular points and find their supremum. Therefore their methods are not applicable here. Shimodaira (2008) considered a multiple scale bootstrap approach to find the corrected ESP values and proved the corrected values are approximately unbiased. It is very similar to those simulation-extrapolation methods which were originally developed to solve measurement error problems (Stefanski and Cook, 1995; Carroll et al., 1996). However, it may require an extensive amount of computing resources for large m.

We propose a novel correction procedure called validation test to handle the bad scenarios in ESP procedure that are caused by singular points. The validation test helps us to decide whether or not a rejection decision given by naïve ESP procedure is convincing in presence of singular points. If the corresponding validation test cannot be rejected, we should be alerted that the rejection decision made by ESP is likely due to the effects from nearby singular points.

Without loss of generosity, we assume the ordered counts $X_1 \leq X_2 \leq \ldots \leq X_m$, and hence permutation p-values $\widehat{p}_1 \leq \widehat{p}_2 \leq \ldots \leq \widehat{p}_m$. Due to the property of BH that deciding rejection on any larger permutation p-values must imply deciding rejection on a smaller permutation p-value. we again develop a step-wise testing method under the same framework of Romano and Wolf (2005), testing from the smallest to largest permutation p-values. Including the first null hypothesis, the more general i-th validation hypothesis

we test is,

$$H_{i,0}^{\text{val}} : p_i^* \in \mathcal{A}_i \text{ and there is an influential singular point nearby}$$
$$H_{i,a}^{\text{val}} : p_i^* \in \mathcal{R}_i \text{ or there is not an influential singular point nearby.}$$

As our testing strategy, we create the empirical null distribution centering at a singular boundary point, and reject the validation null hypothesis if the observed permutation p-value does not seem likely to be drawn from that empirical null distribution.

In particular, we consider the point estimator $\ddot{\boldsymbol{p}} = (\ddot{p}_1, \ddot{p}_2, \ldots, \ddot{p}_m)$, where $\ddot{p}_i = \max(\widehat{p}_i, q \times i/m)$. The point estimator enjoys following properties. First, $\ddot{\boldsymbol{p}}$ is always located at BH decision boundary. This is because the at least one permutation p-value has been rejected by ESP approach, there must be at least one $\widehat{p}_i \leq q \times i/m$. The estimator $\ddot{\boldsymbol{p}}$ could be approximately viewed as MLE of $\boldsymbol{p}^*$ constrained by the closure of acceptance region, if we believe the correlation in the binomial likelihood is negligible. Further more, a $\ddot{\boldsymbol{p}}$ value is singular, if there exist two different indices i_1, i_2 such that $\widehat{p}_{i_1} \leq q \times i_1/m$ and $\widehat{p}_{i_2} \leq q \times i_2/m$. Now we test $H_{1,0}^{\text{val}}$ through testing against point wise null $\ddot{\boldsymbol{p}}$. Since in ESP procedure we already have the empirical distribution of $\widehat{\boldsymbol{p}}^{(b)} = (\widehat{p}_1^{(b)}, \widehat{p}_2^{(b)}, \ldots, \widehat{p}_m^{(b)})$, we are able to approximately obtain the empirical null distribution centering at $(\ddot{p}_1, \ddot{p}_2, \ldots, \ddot{p}_m)$ by mean shifting: the b-th bootstrap sample around the $\ddot{\boldsymbol{p}}$ is generated by $(\widehat{p}_1^{(b)}, \widehat{p}_2^{(b)}, \ldots, \widehat{p}_m^{(b)}) + (\ddot{p}_1 - \widehat{p}_1, \ddot{p}_2 - \widehat{p}_2, \ldots, \ddot{p}_m - \widehat{p}_m) - (r_1, r_2, \ldots, r_m)$. In addition to mean shifting, here $r_1, r_2, \ldots, r_m$ are randomly sampled from $\text{U}[0, \frac{1}{n+1}]$ distribution. These small perturbations will add smoothness to those $\widehat{p}_1^{(b)}, \widehat{p}_2^{(b)}, \ldots, \widehat{p}_m^{(b)}$ because they can only be one of the discrete values in $1/(n+1), 2/(n+1), \ldots, n/(n+1), 1$. Then we use q-value (Storey, 2003) as the validation test statistic. In FDR literature, q-value usually serves as a statistic that measures significance level. A hypothesis is rejected under FDR nominal level q (i.e., 10%) is equivalent to say its corresponding q-value $\leq q$. For the standard BH procedure, the q-value for p-value p_i would be written as

$$q_i = q(p_i) = \min_{t \geq p_i} \widehat{\text{FDR}}(t), \text{ where for any } t, \ \widehat{\text{FDR}}(t) = \frac{mt}{\#\{p_i \leq t\}}, \quad i = 1, 2, \ldots, m. \tag{3.1}$$

We notice that q-values are always sorted in the same order as p-values. We calculate the smallest q-value, $\ddot{q}^{(b)}_{(1)}$, based on the b-th null replicate, $\ddot{\boldsymbol{p}}^{(b)}$, and obtain a validation test p-value by comparing the smallest q-value of $\widehat{\boldsymbol{p}}$ to the collection of $\ddot{q}^{(b)}_{(1)}$. If the validation test p-value is smaller than α, then we reject $\widehat{p}_1$ as the same decision made by the naïve ESP procedure. Otherwise, the rejection by naïve ESP might be unreliable. After we decide whether or not to reject the first validation hypothesis, we move to the next step for testing $\widehat{p}_2$. Details for this general step-wise procedure are presented in Algorithm 2. The validation test will stop at the first time when we fail to reject a validation null. For those tests that are rejected by naïve ESP approach but cannot be rejected by the validation test, we will report them as *acceptance* decisions in output.

3.2.2 The Two-Step Approach

In the previous section, we point out that the naïve ESP procedure may sometimes fail to control the family-wise MCER-I. Now we reexamine the original multiple testing problems:

$$H_{i,0} : p_i^* > \tau^* \qquad \text{versus} \qquad H_{i,1} : p_i^* \leq \tau^*.$$

Because $\{V^{\text{MC-I}} \geq 1\} = \{\text{there exists at least one } H_{i,0} \text{ that is true but rejected}\}$, a rule that controls the FWER of this multiple-testing problem at α will automatically control the family-wise MCER-I at α.

Because τ^* is unknown, testing these hypotheses can be intractable. We propose a two-step procedure that separates the problem into two sub-problems that are more tractable. In the first step, we develop a method that can be used to construct an at least $(1-\beta)$ level one-sided confidence interval $[\tau_l, 1]$ for τ^*, where β is chosen such that $0 < \beta < \alpha$. In the second step, we consider a revised multiple-testing problem, treating τ_l as fixed:

$$\widetilde{H}_{i,0} : p_i^* > \tau_l \qquad \text{vs.} \qquad \widetilde{H}_{i,1} : p_i^* \leq \tau_l, \qquad i = 1, 2, \ldots, m. \tag{3.2}$$

We develop a method to test $\widetilde{H}_{1,0}, \ldots, \widetilde{H}_{m,0}$ that controls the FWER at level $(\alpha - \beta)$. Here each $\widetilde{H}_{i,0}$ is a surrogate for the original hypothesis $H_{i,0}$, and the decision on $H_{i,0}$ is

made by the same decision on $\widetilde{H}_{i,0}$. Then, by the fact that

$$\begin{aligned}\{V^{\text{MC-I}} \geq 1\} &= \{\tau_l > \tau^*, V^{\text{MC-I}} \geq 1\} \cup \{\tau_l \leq \tau^*, V^{\text{MC-I}} \geq 1\} \\ &\subset \{\tau_l > \tau^*\} \cup \{\tau_l \leq \tau^*, \text{ at least one } H_{i,0} \text{ are rejected but } p_i^* > \tau^*\} \\ &\subset \{\tau_l > \tau^*\} \cup \{\tau_l \leq \tau^*, \text{ at least one } \widetilde{H}_{i,0} \text{ are rejected but } p_i^* > \tau_l\},\end{aligned}$$

the family-wise MCER-I $\Pr(V^{\text{MC-I}} \geq 1)$ is always bounded by the sum of $\Pr(\tau_l > \tau^*)$ and the FWER in testing $(\widetilde{H}_{1,0}, \ldots, \widetilde{H}_{m,0})$, which is $\beta + (\alpha - \beta) = \alpha$. As the default choice, we choose $\beta = \alpha/2$, i.e., the total error rate is equally partitioned to the two steps. We show in Appendix B.2 that the equal partition scheme generally yields the highest power.

We also want to point out that, although we focus on controlling the type-I MCER in this work, we can easily extend this two-step algorithm to control the type-II MCER as well. We will present this extension at the end of Appendix B.

Step 1

In step 1, we construct an at least $(1 - \beta)$ level one-sided confidence interval $[\tau_l, 1]$ for τ^*. We find τ_l from a series of cutoffs that are obtained by applying the BH procedure on a series of values that are more conservative than $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_m$, which are permutation p-values $\widehat{p}_i = (X_i + 1)/(n + 1)$ (Phipson and Smyth, 2010). To this aim, we consider the following shrinkage estimator for p_i^* $(i = 1, \ldots, m)$ indexed by a positive tuning parameter c:

$$p_{c,i} = \frac{X_i + c\sqrt{X_i + 1}}{n + c\sqrt{X_i + 1}}. \tag{3.3}$$

The penalty term $\sqrt{X_i + 1}$ in (3.3) is an approximation to $\sqrt{X_i(1 - X_i/n)}$ which is the standard deviation of X_i. We omit $1 - X_i/n$ for ease of computation, since for those $H_{i,0}$ that should be rejected the term should be very close to 1. In addition, we add one to guarantee that the penalty term is always positive. Obviously $p_{c,i}$ is a consistent estimator of p_i^*.

The empirical distribution functions of $(p_1^*, \ldots, p_m^*)$ and $(p_{c,1}, \ldots, p_{c,m})$ are

$$F^*(t) = m^{-1} \sum_{i=1}^{m} \mathbb{I}(p_i^* \leq t)$$
$$F_c(t) = m^{-1} \sum_{i=1}^{m} \mathbb{I}(p_{c,i} \leq t),$$

respectively, for $t \in [0, 1]$. Because $\lim_{c\to\infty} p_{c,i} = 1$, when c is sufficiently large, we expect that the smallest shrinkage estimators are consistently more conservative than the smallest ideal p-values, i.e., $p_{c,(i)} \geq p^*_{(i)}$ for $i = 1, \ldots, l$, where l is the smallest integer such that $p^*_{(l)} \geq q$. We can rewrite this statement in terms of the empirical distribution functions, i.e., $F_c(t) \leq F^*(t)$ for $t \subset [0, q]$. It is known that the BH cutoff can also be obtained by using the empirical distribution function (Storey et al., 2004; Genovese and Wasserman, 2004). Thus the BH cutoffs for $(p_1^*, \ldots, p_m^*)$ and $(p_{c,1}, \ldots, p_{c,m})$, denoted by τ^* and τ_c, can be written as

$$\tau^* = \max\{t : F^*(t) \geq t/q,\ t \in [0, 1]\}$$
$$\tau_c = \max\{t : F_c(t) \geq t/q,\ t \in [0, 1]\}.$$

As Figure 3.2 illustrates, the intersection of either $F^*(t)$ or $F_c(t)$ with the line t/q (if the intersection exists) is always within $[0, q]$, so the BH cutoff is fully determined by the empirical distribution function of p-values on $[0, q]$. If $F_c(t)$ is always no greater than $F^*(t)$ on $[0, q]$ as depicted in Figure 3.2, we must have $\tau_c \leq \tau^*$. Thus we have

$$\Pr(\tau_c \leq \tau^*) \geq \Pr\left(\max_{t\in[0,q]} \{F_c(t) - F^*(t)\} \leq 0\right),$$

where τ_c and $F_c(t)$ are random terms due to $(X_1, \ldots, X_m)$ through $\{p_{c,i}\}$. To construct a $(1 - \beta)$ confidence interval for τ^*, we want to find a tuning parameter c such that

$$\Pr\left(\max_{t\in[0,q]} \{F_c(t) - F^*(t)\} \leq 0\right) \geq 1 - \beta. \tag{3.4}$$

The probability on the left hand side of (3.4) is an increasing functional of c as c increases, because each $p_{c,i}$ increases and thus $F_c(t)$ for a given t decreases as c increases. In order

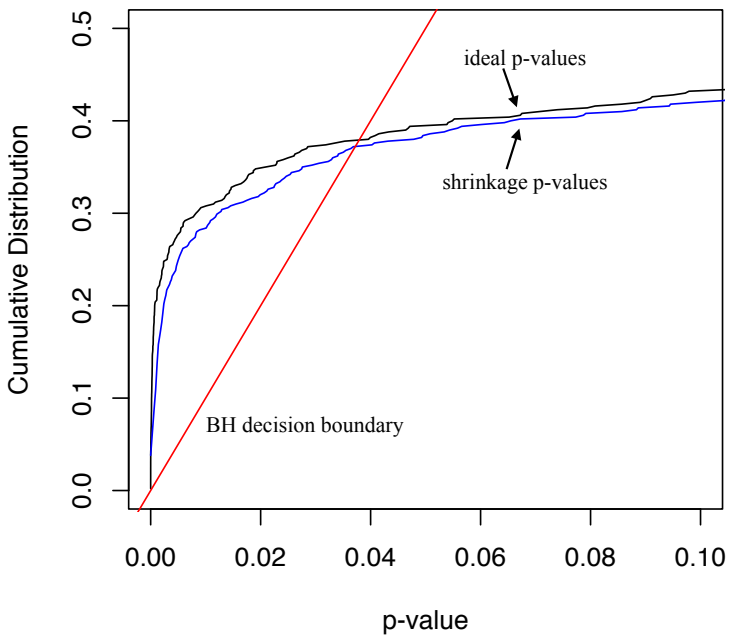

Figure 3.2: An illustrative example of the empirical distribution function of shrinkage p-values F_c versus the empirical distribution function of ideal p-values F^*. Under the standard BH procedure with the FDR nominal level $q = 10\%$, the X-coordinate of the point where BH decision line intersects with F_c is τ_c. The other crossing with F^* corresponds to τ^*, which must be $\geq \tau_c$ in this example. The red line is t/q.

to have the tightest confidence interval for τ^*, we pick the smallest c that satisfies (3.4), and use the corresponding τ_c as the low bound τ_l.

The smallest c among all c that satisfy the inequality (3.4), denoted by c_l, does not have a closed form. Alternatively, we can obtain c_l by applying the bootstrap principle. We obtain the b-th bootstrap sample of $(X_1, \ldots, X_m)$, $(X_1^{(b)}, X_2^{(b)}, \ldots, X_m^{(b)})$ using the same bootstrap sampling scheme in the previous section. The number of bootstrap samples B in this section also needs to be large, but does not have to increase with m. For instance, we can choose $B = 10^4$. Similar to (3.3), we define the shrinkage estimator with the b-th bootstrap sample:

$$p_{c,i}^{(b)} = \frac{X_i^{(b)} + c\sqrt{X_i^{(b)} + 1}}{n + c\sqrt{X_i^{(b)} + 1}},$$

where $i = 1, 2, \ldots, m$. Define the empirical distribution functions

$$F_c^{(b)}(t) = m^{-1} \sum_{i=1}^{m} \mathbb{I}(p_{c,i}^{(b)} \leq t)$$
$$\widehat{F}(t) = m^{-1} \sum_{i=1}^{m} \mathbb{I}(\widehat{p}_i \leq t).$$

Then we obtain c_l as the smallest c that satisfies

$$\frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left(\max_{t \in [0,q]} \left\{ F_c^{(b)}(t) - \widehat{F}(t) \right\} \leq 0 \right) \geq 1 - \beta, \tag{3.5}$$

which is the empirical version of (3.4). The bootstrap reasoning is sketched in Figure 3.3. To this aim, we find, for every $b = 1, \ldots, B$, the smallest c that satisfies $\max_{t \in [0,q]} \left\{ F_c^{(b)}(t) - \widehat{F}(t) \right\} \leq 0$. Then we find c_l as the $(1-\beta)$ quantile of $(c^{(1)}, c^{(2)}, \ldots, c^{(B)})$, which ensures that a proportion $(1 - \beta)$ of $c^{(b)}$s have the indicator function being 1 in (3.5). Finally, we find the low bound of the confidence interval τ_l from c_l. We summarize the entire two-step approach in Algorithm 3.

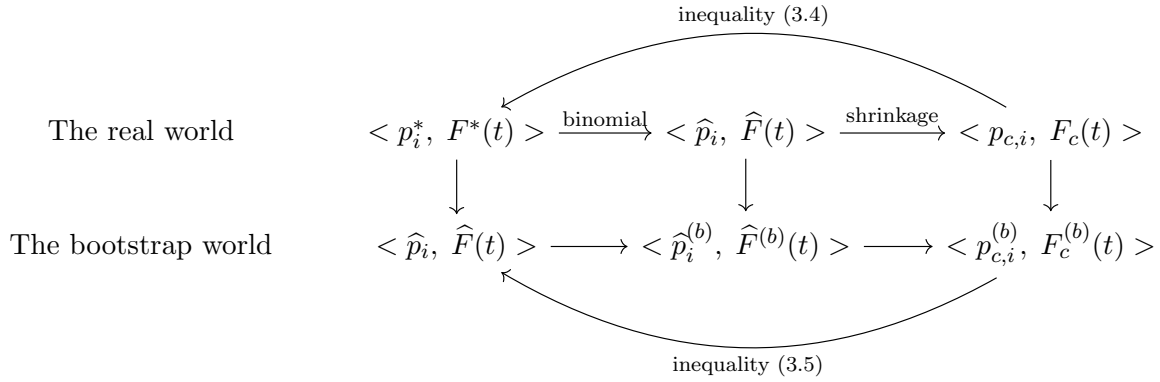

Figure 3.3: Use bootstrap principle to calculate optimal tuning parameter c_l. The bootstrap version of permutation p-value is denoted by $\widehat{p}_i$, and the empirical distribution function is denoted by $\widehat{F}^{(b)}(t)$.

Step 2

Now we present a method to test the hypotheses (3.2) controlling the FWER at level $(\alpha - \beta)$. It is well known that, for testing multiple hypotheses, step-wise procedures such as Holm's are more powerful than single-step procedures such as Bonferroni's. Similar

to the validation test in previous section, we develop a step-wise procedure following the proposal in Romano and Wolf (2005).

We sort $X_1, X_2, \ldots, X_m$ in an ascending order $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(m)}$. We first test the global null, $\widetilde{H}_{(1),0} \cap \widetilde{H}_{(2),0} \cap \ldots \cap \widetilde{H}_{(m),0}$, which means that all p_i^* are greater than τ_l. We use $X_{(1)}$ as the statistic. The asymptotic p-value for this test is given by

$$\Pr\left\{\exists \text{ at least one } \mathrm{Bin}[n, \widetilde{p}_{(i')}] \leq X_{(1)},\ i' = 1, \ldots, m \middle| \widetilde{H}_{(1),0}, \widetilde{H}_{(2),0}, \ldots, \widetilde{H}_{(m),0}\right\},$$

where $\widetilde{p}_{(i)}$ is an estimator of $\max(\tau_l, p_{(i)})$, and $\mathrm{Bin}(n, \theta)$ represents a binomial random variable with the n trials and success rate θ. Followed by the law of the iterated logarithm (Khintchine, 1924), we pick the estimator

$$\widetilde{p}_{(i)} = \begin{cases} \tau_l, & \text{if } \frac{X_{(i)} - n\tau_l}{\sqrt{\tau_l(1-\tau_l)}} \leq \sqrt{2n\log(\log(n))}, \\ \frac{X_{(i)}}{n}, & \text{otherwise,} \end{cases}$$

which was proposed in the work of Hansen (2005) and Hsu et al. (2010). According to the Bonferroni inequality, the p-value is less than

$$\sum_{i'=1}^{m} \Pr\left\{\mathrm{Bin}[n, \widetilde{p}_{(i')}] \leq X_{(1)}\right\}. \tag{3.6}$$

Therefore it works under any dependence structure among the m original tests. If the term in (3.6) is less than $(\alpha - \beta)$, we reject $\widetilde{H}_{(1),0}$ and move to $X_{(2)}$; otherwise, we stop testing any larger $X_{(i)}$ and declare that none of the null hypotheses should be rejected. In general, when it comes to $X_{(i)}$ for $i > 1$, we use $X_{(i)}$ to test the joint null $\widetilde{H}_{(i),0} \cap \ldots \cap \widetilde{H}_{(m),0}$. If

$$\sum_{i'=i}^{m} \Pr\left\{\mathrm{Bin}[n, \widetilde{p}_{(i')}] \leq X_{(i)}\right\}$$

is less than $(\alpha - \beta)$, we reject $\widetilde{H}_{(i),0}$ and move to $X_{(i+1)}$; otherwise, we stop further testing and accept the remaining hypotheses. Because of the closure principle (Marcus et al., 1976), this step-wise procedure controls the FWER in tests (3.2) by $(\alpha - \beta)$. More technical presentation can be found in Algorithm 3.

Algorithm 3 The two-step approach for controlling type-I MCER

Input: the matrix $\boldsymbol{I}$, nominal FDR rate q, nominal family-wise MCER-I α

Step 1: Constructing the $(1-\beta)$ level confidence interval $[\tau_l, 1]$ for τ^*.

(1.1) For each $b \in (1, 2, \ldots, B)$, generate a bootstrap sample $(X_1^{(b)}, X_2^{(b)}, \ldots, X_m^{(b)})$ based on $\boldsymbol{I}^{(b)}$.

(1.2) Find $c^{(b)}$ to be the smallest c that satisfies $\max_{t\in[0,q]} \left\{ F_c^{(b)}(t) - \widehat{F}(t) \right\} \leq 0$.

(1.3) Find c_l as the $(1-\beta)$ quantile of $(c^{(1)}, c^{(2)}, \ldots, c^{(B)})$.

(1.4) Find τ_l by applying the BH procedure on shrinkage estimators indexed by c_l.

Step 2: Testing multiple hypotheses (3.2) given τ_l.

(2.1) For $i = 1, 2, \ldots, m$, if $\sum_{i'=i}^{m} \Pr\left\{ \mathrm{Bin}[n, \widetilde{p}_{(i')}] \leq X_{(i)} \right\} \leq \alpha - \beta$, reject $\widetilde{H}_{(i),0}$, update $i = i+1$, and rerun (2.1); otherwise, accept all remaining hypotheses and stop.

(2.2) If $i > m$ then stop.

(2.3) Whenever $\widetilde{H}_{i,0}$ is rejected, $H_{i,0}$ is rejected.

Output: a list of rejected hypotheses.

3.3 Simulation Studies

3.3.1 Setup

We evaluated the following four methods: (i) the two-step procedure, (ii) the naïve ESP procedure only, (iii) applying the ESP procedure and the validation test, and (iv) a variant of the methods proposed by Gandy and Hahn (2016) to ensure a fair comparison to our methods with fixed n. We call this variant GH-fixed method. It differs from the original method in the following aspects. First, since we only wish to control the error rates of making non-reproducible rejection, a one-sided confidence interval $(0, p_i^{u'}]$ at $(1-\alpha/m)$ level for each p_i^* would be sufficient. Second, Gandy and Hahn (2016) suggested to use the Robbins-Lai confidence interval (Lai, 1976) to maintain the exact coverage even when different tests stop after different number of permutations due to early stopping in the sequential procedure. However, the Robbins-Lai interval is usually too wide, particularly wider than the intervals using fixed sample sizes (i.e., non-sequential) (Coe and Tamhane, 1993), and hence leads to loss of power. Also, the Robbins-Lai interval cannot be one-sided. In our implementation, we used Wilson confidence interval (Wilson, 1927), which is robust and accurate under very general situations, although not applicable to the sequential problem, and is recommended by Brown et al. (2001,

2002). Third, we apply the standard BH on the upper limits of confidence intervals, $p_1^{u'}, p_2^{u'}, \ldots, p_m^{u'}$, obtain the rejection set $\mathcal{R}' \subset \{1, 2, \ldots, m\}$, and decide the hypotheses in set $\mathcal{R}'$ to be rejected and the remaining hypotheses to be accepted; there would not be undecided hypotheses. This way, we also obtain an algorithm for a fixed n. One can use similar arguments to those in Gandy and Hahn (2014, 2016) to prove that this procedure controls the family-wise MCER-I defined for our method.

We considered $m = 1000$ tests. In the first set of experiments, we evaluated performances of the methods under a general scenario in which ideal p-values were sampled from a parametric distribution. Specifically, we assumed that a proportion π_0 of m tests are under the null hypothesis and the remaining tests are under the alternative. We set $\pi_0 = 80\%$. Different choice of π_0 will not change the results of simulations. The p-values for tests under the null were sampled independently from the uniform distribution U$[0, 1]$. Those under the alternative were sampled independently from the Gaussian right-tailed probability model, $p_i^* = 1 - \Phi(Z_i)$, where $Z_i \sim \mathrm{N}(\beta, 1)$ and Φ is the standard normal cumulative distribution function. Here β represents the strength of association, and was chosen to be 1.5, 2 and 2.5. A larger value of β implies a greater chance for a non-null hypothesis to be rejected. Given a fixed set of m, π_0 and β, we sampled 100 different realizations of $(p_1^*, p_2^*, \ldots, p_m^*)$. We generated the rejection table $\boldsymbol{I}$ by drawing Bernoulli samples with success probabilities $(p_1^*, p_2^*, \ldots, p_m^*)$ instead of generating actual permutations. We varied the number of permutations n from 5,000 to 160,000. The nominal level of family-wise MCER-I α was set to 10%; the nominal FDR q was set to 10%.

In the second set of experiments, ideal p-values were chosen to be close to those singular points at BH decision boundary. It is easy to verify that for any integer $1 < k \leq m$, $\ddot{\mathbf{p}}_{s_1} = (q/m,\ 2q/m,\ \cdots,\ kq/m,\ (k+1)/m,\ (k+2)/m,\ \cdots,\ 1)$ is a type-I singular point and $\ddot{\mathbf{p}}_{s_2} = (kq/m,\ kq/m,\ \cdots, kq/m,\ (k+1)/m,\ (k+2)/m,\ \cdots, 1)$ is a type-II singular point. Here we point out that k is the number of coordinates that are right at the BH check points in either $\ddot{\mathbf{p}}_{s_1}$ or $\ddot{\mathbf{p}}_{s_2}$. In the set of experiment k is chosen from $10, 80$ and 150 respectively. Then ideal p-values $\mathbf{p}^*$ were given by $(1+\epsilon)\ddot{\mathbf{p}}_{s_1} \wedge 1$ which could be viewed as a subtle shift from a type-I singular point; or $(1+\epsilon)\ddot{\mathbf{p}}_{s_2} \wedge 1$ which is shifted from a type-II singular point. We chose ϵ to be $10^{-1}, 10^{-2}, 10^{-3}$. A smaller ϵ implies a shorter

distance to the singular points.

We used two metrics to evaluate the performance of each method. One is the empirical family-wise MCER-I and the other one is the average proportion of "ground truth" rejections that are rejected by each method, i.e.,

$$\frac{\sum_{i=1}^{m} \mathbb{I}(\widehat{D}_i = \textit{rejection}, D_i^* = \textit{rejection})}{1 \vee \sum_{i=1}^{m} \mathbb{I}(D_i^* = \textit{rejection})}.$$

We refer to the second metric as detection sensitivity. Both metrics were based on 1,000 simulation replicates.

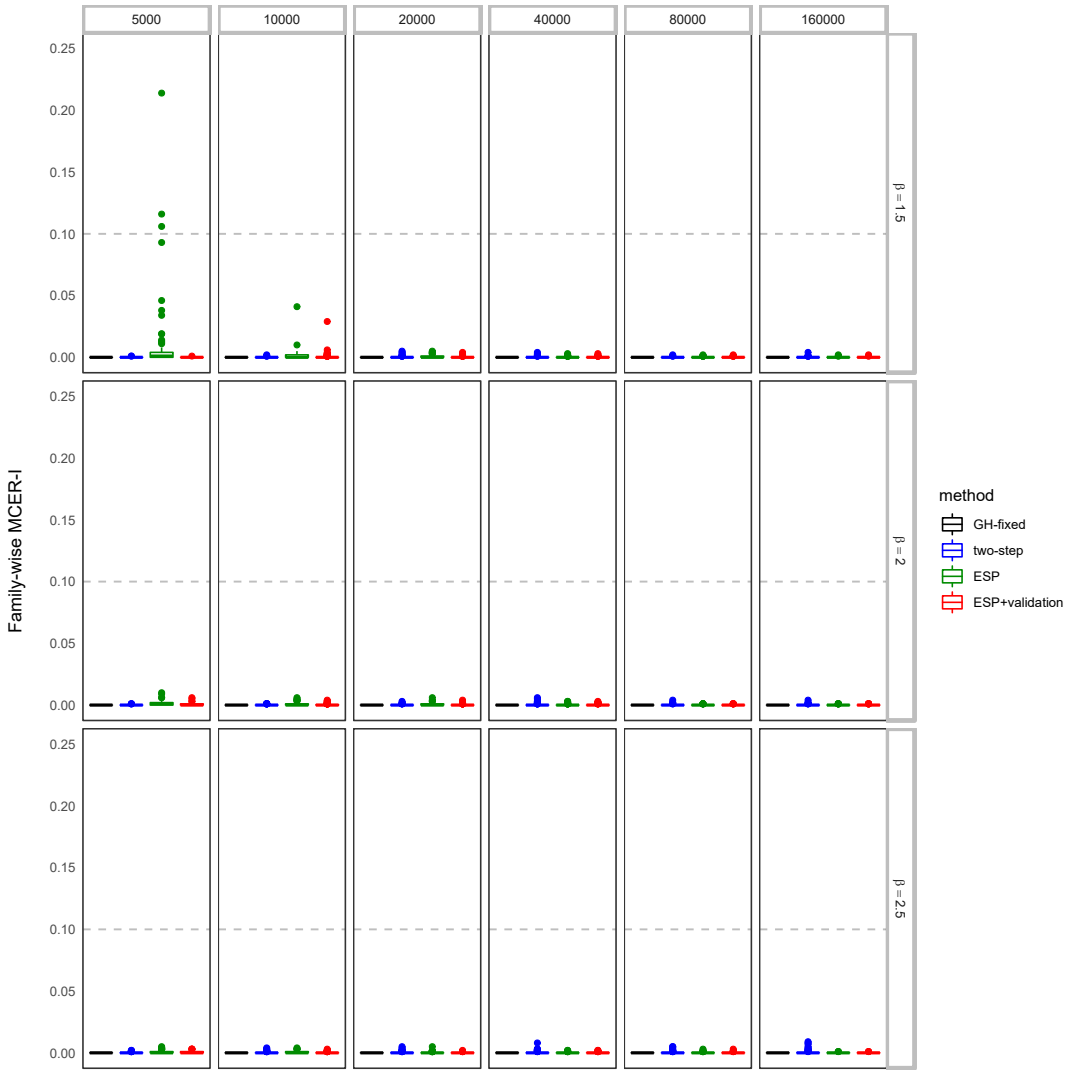

Figure 3.4: The empirical family-wise MCER-I under different number of permutations (from 5k to 160k). The non-null ideal p-values were simulated from the Gaussian right-tailed model with $\beta = 1.5, 2.0$ and 2.5.

3.3.2 Simulation Results

The results of the first set of simulations are shown in Figure 3.4 and 3.5. The empirical family-wise MCER-I and detection sensitivity are shown in boxplots where the top and bottom of the box are 75% and 25% quantiles and the middle band indicates the median. Figure 3.4 shows that in all scenarios the two-step approach, ESP procedure with validation test, and the GH-fixed method controlled family-wise MCER-I below the nominal level. The ESP procedure without validation cannot always control the error rate. In particular, there is inflation of family-wise MCER-I in three realizations when $\beta = 1.5$ and $n = 5,000$. The error rate with the GH-fixed method is always the most conservative among all methods. That method turns out to be the least powerful one. Compared to that method, both two-step and ESP methods demonstrated substantial improvement in detection sensitivity. The ESP approach seems usually the most powerful one, although it cannot always control error rates. The ESP procedure with validation test is usually the second most powerful method, except the cases in which the number of replicates is small ($n = 5,000$) and the two-step method gains more power there.

The results of the second set of simulations are shown in Table 3.1, 3.2 and 3.3. The two-step method and GH-fixed method always control family-wise MCER-I by 10% in all scenarios. The ESP approach shows severe error rate inflation when the ideal p-values are close to type-I singular points (corresponds to $\epsilon = 10^{-2}, 10^{-3}$). The worst scenarios are shown in the Table 3.2 and Table 3.3, where 80 or 150 coordinates of ideal p-values are right above the BH check points: the empirical family-wise MCER-I can reach 100% and on average we observe a significant amount of type-I MC errors. The validation test is helpful to eliminate most of the inflated errors. In another experiment where the ideal p-values are close to type-II singular points, none error is observed in all cases. These results suggest that type-I singular point is an important factor that causes inflated error rates with naïve ESP procedure.

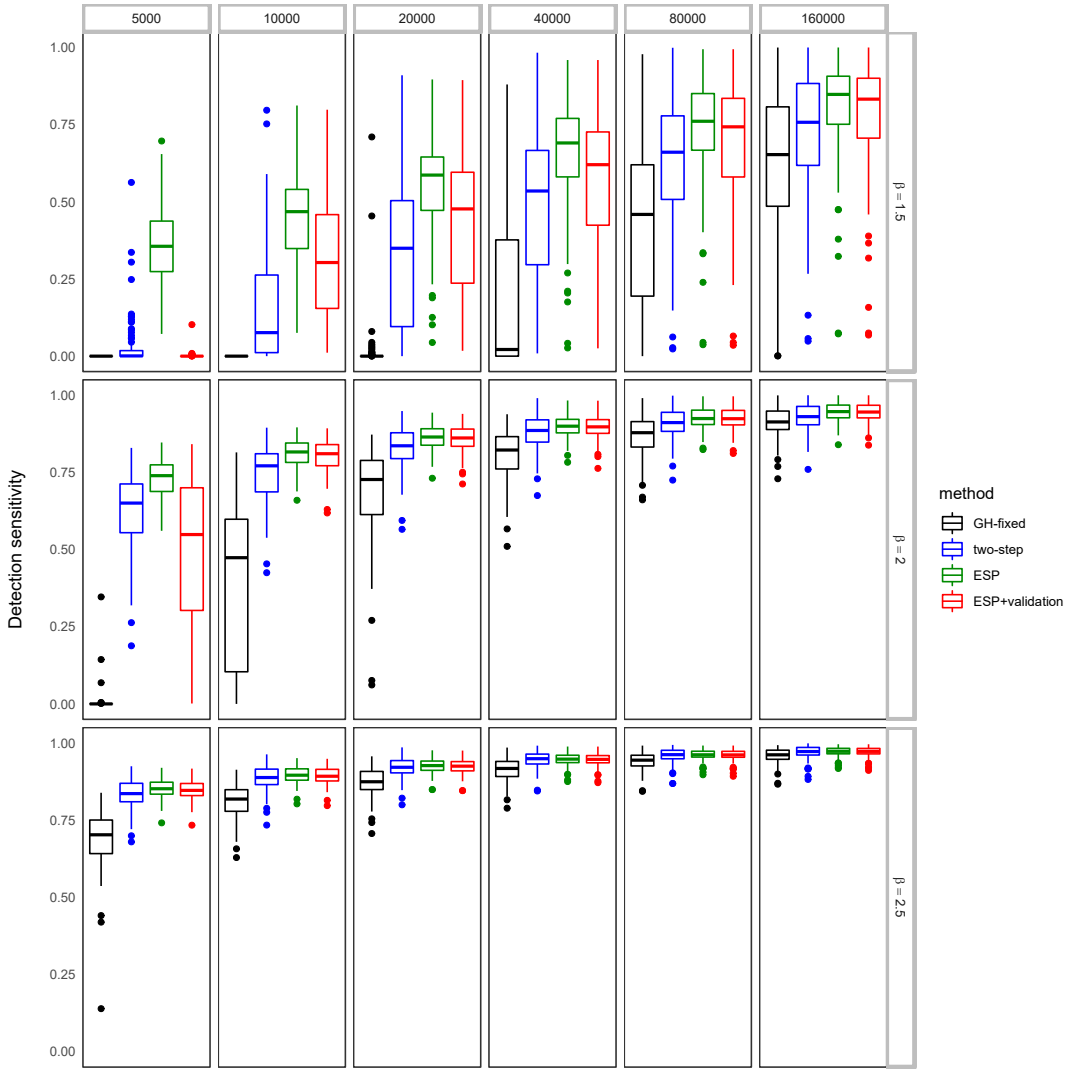

Figure 3.5: Detection sensitivity in all 100 different realizations under different number of permutations (from 5k to 160k). The non-null ideal p-values were simulated from the Gaussian right-tailed model with $\beta = 1.5, 2.0$ and 2.5.

Table 3.1: The family-wise MCER-I, and the average number of type-I MC errors when the ideal p-values are close to type-I singular points and $k = 10$ coordinates are close to the BH check points.

ϵ	#replicates	two-step		ESP		ESP+validation		GH-fixed	
		MCER-I	$\mathbb{E}V^{\text{MC-I}}$	MCER-I	$\mathbb{E}V^{\text{MC-I}}$	MCER-I	$\mathbb{E}V^{\text{MC-I}}$	MCER-I	$\mathbb{E}V^{\text{MC-I}}$
10^{-1}	10,000	0%	0	43.1%	0.495	5.6%	0.099	0%	0
	40,000	1.5%	0.017	4.7%	0.06	2.9%	0.03	0%	0
	160,000	0.2%	0.002	1.0%	0.011	0.9%	0.009	0%	0
10^{-2}	10,000	0%	0	50.4%	0.609	7.3%	0.132	0%	0
	40,000	3.3%	0.038	18.6%	0.334	6.7%	0.092	0%	0
	160,000	1.1%	0.014	32.7%	0.744	8.8%	0.14	0%	0
10^{-3}	10,000	0%	0	48.9%	0.589	8.0%	0.137	0%	0
	40,000	2.1%	0.027	21.7%	0.428	6.8%	0.083	0%	0
	160,000	2.3%	0.056	36.0%	0.918	9.2%	0.157	0%	0

Table 3.2: The family-wise MCER-I, and the average number of type-I MC errors when the ideal p-values are close to type-I singular points and $k = 80$ coordinates are close to the BH check points.

ϵ	#replicates	two-step		ESP		ESP+validation		GH-fixed	
		MCER-I	$\mathbb{E}V^{\text{MC-I}}$	MCER-I	$\mathbb{E}V^{\text{MC-I}}$	MCER-I	$\mathbb{E}V^{\text{MC-I}}$	MCER-I	$\mathbb{E}V^{\text{MC-I}}$
10^{-1}	10,000	0%	0	48.5%	0.801	5.2%	0.094	0%	0
	40,000	1.2%	0.014	10.0%	0.179	4.0%	0.051	0%	0
	160,000	0.9%	0.010	1.8%	0.026	0.8%	0.011	0%	0
10^{-2}	10,000	0%	0	96.4%	14.667	7.8%	0.148	0%	0
	40,000	1.3%	0.017	100%	28.24	9.6%	0.143	0%	0
	160,000	2.3%	0.041	100%	38.899	12.5%	0.237	0%	0
10^{-3}	10,000	0.1%	0.004	98.2%	19.075	7.5%	0.146	0%	0
	40,000	3.1%	0.037	100%	37.437	11.0%	0.188	0%	0
	160,000	1.8%	0.038	100%	52.671	14.1%	0.284	0%	0

Table 3.3: The family-wise MCER-I, and the average number of type-I MC errors when the ideal p-values are close to type-I singular points and $k = 150$ coordinates are close to the BH check points.

ϵ	#replicates	two-step		ESP		ESP+validation		GH-fixed	
		MCER-I	$\mathbb{E}V^{\text{MC-I}}$	MCER-I	$\mathbb{E}V^{\text{MC-I}}$	MCER-I	$\mathbb{E}V^{\text{MC-I}}$	MCER-I	$\mathbb{E}V^{\text{MC-I}}$
10^{-1}	10,000	0%	0	48.6%	0.816	5.2%	0.094	0%	0
	40,000	0.9%	0.010	7.8%	0.131	3.4%	0.04	0%	0
	160,000	0.5%	0.007	1.8%	0.022	1.2%	0.014	0%	0
10^{-2}	10,000	0%	0	99.8%	40.467	7.8%	0.148	0%	0
	40,000	2.5%	0.030	100%	61.778	11.4%	0.173	0%	0
	160,000	2.4%	0.042	100%	74.328	13.8%	0.259	0%	0
10^{-3}	10,000	0.1%	0.004	100%	55.498	6.6%	0.12	0%	0
	40,000	2.1%	0.025	100%	89.352	10.6%	0.159	0%	0
	160,000	2.6%	0.073	100%	112.19	12.2%	0.252	0%	0

3.4 Application to Prostate Cancer Data

We considered a benchmark dataset (Singh et al., 2002) which contains microarray gene expression data for 6,033 genes and 102 subjects (52 prostate cancer patients and 50 healthy controls); more details of this dataset can be found in Efron (2012). The goal is to detect genes that are differentially expressed between prostate cancer patients and healthy controls. To this end, we adopted the two-sample t-statistic with equal variance for each gene. We calculated the t-statistic for the observed dataset and n permutation datasets, and obtained the matrix $\boldsymbol{I}$ as input for downstream analysis.

We applied the naïve, GH-fixed, and two-step methods. The ESP method was not compared here because it took even more run-time than permutation sampling. For the naïve method, we calculated the permutation p-value for each gene and applied the standard BH procedure with nominal FDR = 10%. This method does not make any effort to control type-I MCER. For the two-step method, the nominal family-wise MCER-I was set to 10% and the nominal FDR was also set to 10%. For comparison purposes, we obtained the results for 10 runs.

In each run, the two-step procedure yields more rejections than the GH-fixed method. For each experiment, among the detected genes by the two-step method, the chance that there exists a detection which would not be detected if we run the permutation for an infinite number is less than 10%. Note that each list of detection is just a subset of the ideal list of discoveries.

The two-step method costs about 20 minutes (not including permutation) for $n = 1,000,000$. However, bootstrap resampling in step 1 is the most computationally intensive part in our algorithm. The run-time of bootstrap resampling can be significantly reduced by implementing our algorithm under a parallel computing framework. Our program costs about 2.5 minutes when we use 10 cores for execution. The run-time can be further reduced with more cores.

Table 3.4: The number of detected genes that are determined to be differentially expressed. $\alpha = 10\%$.

	Run	1	2	3	4	5	6	7	8	9	10
$n = 100,000$											
	Naïve	64	62	60	63	63	60	61	59	60	62
	Two-step	51	52	52	55	52	54	50	49	51	49
	GH-fixed	0	0	0	0	20	0	0	0	0	19
$n = 500,000$											
	Naïve	62	62	62	62	61	60	61	60	60	62
	Two-step	58	58	58	57	57	58	57	57	58	57
	GH-fixed	53	56	56	51	52	52	56	54	51	53
$n = 1,000,000$											
	Naïve	62	62	61	62	62	60	62	59	60	62
	Two-step	58	58	57	58	57	58	58	58	58	57
	GH-fixed	57	57	57	57	57	58	57	58	57	57

3.5 Discussion

In this work, we considered the problem of controlling type-I MCER on multiple resampling-based tests. We focused on a common situation in application, where the number of permutations is a fixed number n (i.e., $= 100,000$). However, in many real studies, the n that a practitioner picks may not be sufficiently large to account for the variability in decisions by testing a large scale of hypotheses. It is known that the computational cost of permutations increases with the number of tests. For instance, it is usually infeasible to generate millions of permutations on whole genome sequencing data. For some applications, the procedure to obtain permutation statistics (e.g., matrix decomposition) may be time-consuming, so it may not be possible to run millions of permutations. With a moderate size of n, it is likely that some rejections given by the naïve methods should be accepted on ideal p-values, which brings us type-I MC error. Our methods are designed to avoid those non-reproducible rejections and try to confirm as many true rejections as possible.

In addition to multiple permutation tests with a fixed n, it is known that sequential algorithms can also be useful in reducing the overall computational costs of multiple permutation tests. We will further investigate the problem related to sequential multiple resampling-based tests in the next chapter.

Chapter 4

Sequential Resampling-Based Multiple Testing Procedure That Controls Monte Carlo Error Rate

4.1 Introduction

In this chapter we propose a class of methods for sequential resampling-based multiple testing called MCTGS (Monte Carlo Tests with Group Sequential approaches). In Section 4.2.1, we point out the connection between the Robbins-Lai interval and a well-known sequential testing procedure. By virtue of the duality principle between confidence set and hypothesis testing, we develop novel sequential confidence intervals in Section 4.2.2. In Section 4.2.3, we show that the Bonferroni correction to construct simultaneous confidence intervals is not necessary, and propose an alternative step-wise procedure that enhances power. Our methods can be shown to control family-wise MCER. In Section 4.3, we demonstrate improved power with our methods through synthetic data. In Section 4.4, we revisit the gene expression data from Section 3.4 with the proposed sequential resampling-based testing approaches.

4.2 Methods

4.2.1 Duality Between Sequential Testing and Confidence Set

It is well known that there is duality (Lehmann and Romano, 2006) between confidence set and hypothesis testing. The follow Lemma 4.1 shows that, for any $\alpha \in (0,1)$, a confidence set (or region) that provides at least $(1-\alpha)$ coverage probability can be derived from the acceptance region of a hypothesis test that controls type-I error by α, regardless of sequential or non-sequential procedures.

Lemma 4.1. Suppose that for every value $p \in [0,1]$, there is a test at level α_i of the hypothesis $H_0 : p_i^* = p$. The observed data are denoted by X. Denote the acceptance region of the test by $A(p) = \{X|\ H_{i,0} : p_i^* = p \text{ cannot be rejected}\}$. Then the acceptance region inversion $\mathcal{A}(X) = \{p|\ X \in A(p)\}$ is a $(1-\alpha_i)$ level confidence set for p_i^*.

Proof. First we know that type-I error $\Pr(X \notin A(p_i^*)|p_i^*) \leq \alpha_i$ and then equivalently $\Pr(X \in A(p_i^*)|p_i^*) \geq (1-\alpha_i)$ when p_i^* is used to sample data X. Based on definitions, $p_i^* \in \mathcal{A}(X)$ is equivalent to $X \in A(p_i^*)$. Hence $\Pr(p_i^* \in \mathcal{A}(X)|p_i^*) = \Pr(X \in A(p_i^*)|p_i^*) \geq$

$(1-\alpha_i)$, which implies that $\mathcal{A}(X)$ satisfies the definition of a $(1-\alpha_i)$ level confidence set. □

In this section we show that the Robbins-Lai interval described by (1.5) is closely connected to the mixture sequential probability ratio test (mSPRT) which was first proposed by Wald (1945). The mSPRT is known as a benchmark procedure in sequential hypothesis testing, which has wide applications such as industrial process control and continuous surveillance. Robbins and Siegmund (1972) further showed that mSPRT is a power-1 testing procedure as long as sufficient samples are collected. Suppose we consider the i-th test: $H_0 : p_i^* = p$ v.s $H_a : p_i^* \neq p$. For any $p' \neq p$ and at the time of collecting j permutation samples, we can define a likelihood ratio between the alternative and null hypotheses:

$$\Lambda_{i,k}[p' : p] = \frac{p'^{X_{i,k}}(1-p')^{k-X_{i,k}}}{p^{X_{i,k}}(1-p)^{j-X_{i,k}}}.$$

Recall that in Section 1.3 we defined $X_{i,k} = \sum_{k'=1}^{k} \mathbb{I}(T_{i,k'} > t_i)$ which is the number of permuted statistics greater than observed statistic within the first k permutations. In mSPRT the apriori mixing distribution for p' under alternative hypothesis can be picked arbitrarily. When the mixing distribution is U$[0,1]$, the integrated likelihood ratio can be written as

$$\tilde{\Lambda}_{i,k}(p) = \int_{[0,1]} \Lambda_{i,k}[p' : p] \times 1 \, dp' = \int_{[0,1]} \frac{p'^{X_{i,k}}(1-p')^{k-X_{i,k}}}{p^{X_{i,k}}(1-p)^{k-X_{i,k}}} dp' = \frac{1}{(k+1)\binom{k}{X_{i,k}} p^x (1-p)^{k-X_{i,k}}}.$$

The denominator in this formula is exactly the same as the left hand side of equation (1.4). Similar to other mSPRT procedures, the integrated likelihood ratio $\tilde{\Lambda}_{i,k}(p)$ is a martingale with respect to filter $\sigma(\cup\{X_{i,k}, k\})$, and under the null condition marginal expectation $\mathbb{E}\tilde{\Lambda}_{i,k}(p) = 1$ holds for every k. The Doob's martingale inequality (Doob, 1953) implies that under the null $(p_i^* = p)$,

$$\Pr(\sup_k \tilde{\Lambda}_{i,k}(p) \geq \frac{1}{\alpha_i}\}) \leq \frac{1}{1/\alpha_i} = \alpha_i.$$

The acceptance region inversion of the test above is

$$\begin{aligned}\mathcal{A}_i^{\mathrm{rl}} &= \{\sup_k \tilde{\Lambda}_{i,k}(p) \geq \frac{1}{\alpha_i}\} \\ &= \{\tilde{\Lambda}_{i,1}(p) < \frac{1}{\alpha_i}\} \cap \{\tilde{\Lambda}_{i,2}(p) < \frac{1}{\alpha_i}\} \cap \cdots \cap \{\tilde{\Lambda}_{i,k}(p) < \frac{1}{\alpha_i}\} \cdots .\end{aligned}$$

With the Bonferroni correction approach for simultaneous error control, we choose all $\alpha_i = \alpha/m$. It is easy to notice that $\mathcal{I}_{i,k}$ defined in (1.5) equals $\{\tilde{\Lambda}_{i,k}(p) < \frac{1}{\alpha_i}\}$, which further implies that the Robbins-Lai interval presented in Section 1.3, is a special case of acceptance region inversion $\mathcal{A}_i^{\mathrm{rl}}$.

4.2.2 Sequential Confidence Intervals Based on Group Sequential Approaches

In this section, we attempt to improve the sequential algorithm of Gandy and Hahn (2016) by replacing the Robbins-Lai interval with another type of interval that can be obtained using the duality principle but with a more powerful sequential testing procedure. Although one can gain more power by replacing the non-informative mixing distribution U[0, 1] with an optimal mixing distribution, little information is known about the distribution of p-values under the alternative hypotheses. In addition to the framework of mSPRT, the other natural choice is to consider the testing procedures based on group sequential approaches. These sequential methods have been developed for decades (Armitage, 1958; Pocock, 1977; O'Brien and Fleming, 1979; Gordon Lan and DeMets, 1983), mainly for decision making in clinical trial. We want to develop confidence intervals based on group sequential tests because some of them can be more powerful than mSPRT. Suppose the total N_{max} permutation samples can be partitioned to G groups. Each group consists of N_{max}/G samples. The first group corresponds to the first N_{max}/G samples; the second group are the $(N_{max}/G+1)$-th, $(N_{max}/G+2)$-th, $\cdots$, and $2N_{max}/G$-th samples; $\cdots$; the last group corresponds to the last N_{max}/G samples. For the i-th test, we use

$$Z_{i,j} = \sum_{k=1}^{N_{max}/G} \mathbb{I}(T_{i,\, k+(j-1)N_{max}/G} > t_i) \qquad (j = 1, 2, \cdots, G) \tag{4.1}$$

to denote the number of permuted statistics that are more extreme than the observed statistic within the j-th group. We know that $Z_{i,1}, Z_{i,2}, \cdots, Z_{i,G}$ are independent and follow $\text{Bin}(N_{max}/G, p_i^*)$ distribution. When the number of samples within a group, N_{max}/G, is sufficiently large, by central limit theorem $Z_{i,j}$ marginally converges to a Gaussian distribution. Similar to existing group sequential methodologies (Pocock, 1977; O'Brien and Fleming, 1979) for Gaussian outcome, we consider the standardized cumulative summation:

$$S_{i,j}(p) = \frac{1}{\sqrt{N_{max} \times p \times (1-p)}} \sum_{l=1}^{j} (Z_{i,l} - N_{max}/G \times p) \qquad (j = 1, 2, \cdots, G), \tag{4.2}$$

such that under null condition $p = p_i^*$, the last term $S_{i,G}(p_i^*)$ follows $\text{N}(0, 1)$. Moreover,

$$\text{Var}(S_{i,j}(p_i^*)) = \frac{j}{G}, \quad \text{and for } j_1 \neq j_2, \ \text{Cov}(S_{i,j_1}(p_i^*), S_{i,j_2}(p_i^*)) = \min(\frac{j_1}{G}, \frac{j_2}{G}).$$

The dependency structure among G statistics is known. Note that this null condition is used to derive two-sided confidence sets. Otherwise, the null condition $p \geq p_i^*$ or $p \leq p_i^*$ leads to one-sided sets. Due to the property of independent increment in $S_{i,j}(p)$ and the asymptotic normality we mentioned above, the joint distribution of $(S_{i,1}(p_i^*), S_{i,2}(p_i^*), \cdots, S_{i,G}(p_i^*))$ converges to $(B(1/G), B(2/G), \cdots, B(1))$, where $B(\cdot)$ represents a standard Brownian motion. To control the type-I error of group sequential test by α_i, we hope to find a series of positive thresholds $c_1, c_2, \cdots, c_G$ such that:

$$\begin{aligned} &\Pr(\exists\ j \in \{1, 2, \cdots, G\} \text{ such that } \ |S_{i,j}(p_i^*)| \geq c_j) \\ &\approx \Pr(\exists\ j \in \{1, 2, \cdots, G\} \text{ such that } \ |B(j/G)| \geq c_j) \leq \alpha_i. \end{aligned} \tag{4.3}$$

There are many possible choices of c_j that satisfy this inequality. Each set of c_j uniquely defines a group sequential test. However, the simplest set of thresholds would be $c_1 = c_2 = \cdots = c_G = c$, which can be viewed as an approximation of the decision rules

proposed in O'Brien and Fleming (1979). Then we rewrite

$$\begin{aligned}
&\Pr(\exists\, j \in \{1, 2, \cdots, G\} \text{ such that } |B(j/G)| \geq c) \\
&\leq \Pr(\exists\, j \text{ such that } B(j/G) \geq c) + \Pr(\exists\, j \text{ such that } B(j/G) \leq -c) \\
&= 2\Pr(\exists\, j \text{ such that } B(j/G) \geq c) \\
&\leq 2\Pr(\exists t \in (0, 1] \text{ such that } B(t) \geq c) \\
&= 2\Pr(\sup_{t\in(0,1]} B(t) \geq c) = 4\Pr(B(1) \geq c).
\end{aligned}$$

The last equality is known as the reflection principle (Durrett, 2019) of Brownian motion. To control this probability by α_i, we can easily calculate the constant thresholds

$$c_j = c = \Phi^{-1}(1 - \alpha_i/4), \tag{4.4}$$

because $B(1) \sim \mathrm{N}(0, 1)$. In simulations, we will show that this set of thresholds achieve high power. For the purpose of comparison, we also consider another set of thresholds which satisfy

$$c_j = \sqrt{\frac{j+1}{G}\left[2\log(\frac{1}{\alpha_i}) + \log(j+1)\right]}. \tag{4.5}$$

It is known that for standard Brownian motion, both

$$\exp(\theta B_t - \frac{1}{2}\theta^2 t) \text{ and } \int_\theta \exp(\theta B_t - \frac{1}{2}\theta^2 t)\pi(\theta)d\theta$$

are martingales with respect to the filter $\sigma(\{B(s) : 0 \leq s \leq t\})$, where θ is a parameter and π is an arbitrary mixing distribution for parameter θ. Followed by the results of Robbins (1970) and 's martingale inequality (Doob, 1953), it can be shown that the inequality (4.3) still holds for standard Brownian motion under the thresholds in (4.5). Approximately these thresholds have an parabolic shape. Because the log term $\log(j+1)$ is less dominant compared to the constant $2\log(1/\alpha_i)$, c_j in (4.5) increases in an order of $\sqrt{j}$. These thresholds mimic the behaviours of the group sequential approach by Pocock (1977), but differ from the the original proposal. In Pocock (1977), $c_j/\sqrt{j}$ must be a constant but there is no closed-form solution to this constant. When it comes to the

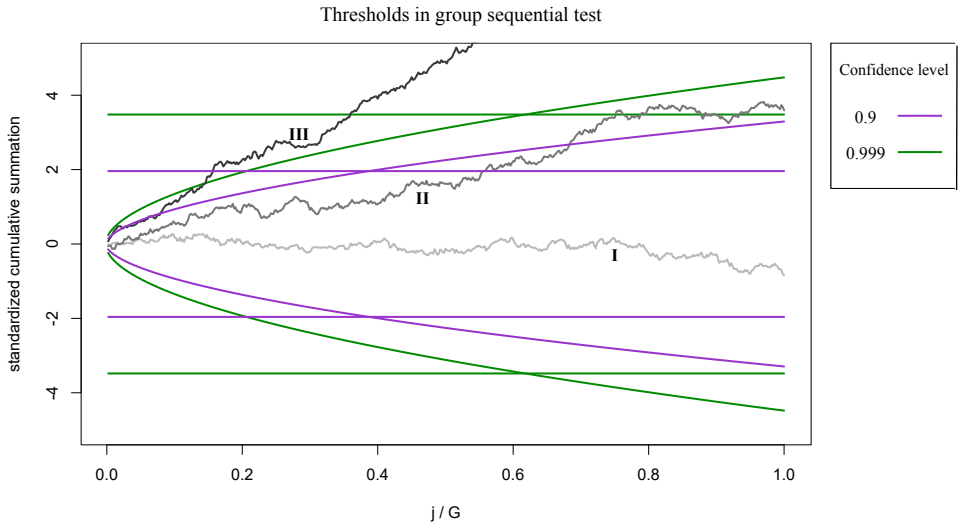

Figure 4.1: Standardized cumulative summation under three parameters $\{S_{i,j}(p_1) : j = 1, 2, \cdots G\}$, $\{S_{i,j}(p_2) : j = 1, 2, \cdots G\}$, $\{S_{i,j}(p_3) : j = 1, 2, \cdots G\}$ are denoted by paths I II III respectively. Only p_1 is under the null condition (equals to the truth from sampling distribution), while $0 < |p_2 - p_1| < |p_3 - p_1|$. For the first time when sample path meets these thresholds, we reach a conclusion that the values (e.g., p_2 and p_3) are outside the confidence sets. If the difference from null condition is sufficiently large (i.e., path III), the sample path will first exceed the 'parabolic' thresholds and later meet with the 'flat' thresholds. However, if difference is small (i.e., path II), using the 'flat' thresholds can be more efficient.

end of sequential permutation tests, these thresholds in (4.5) should be larger than the thresholds in (4.4). Figure 4.1 gives an illustrative example for two types of thresholds.

So far we have found an α_i level group sequential test for testing null hypothesis $H_{i,0} : p_i^* = p$ using the sequential test statistics $S_{i,j}(p),\ j = 1, 2, \cdots, G$. We reject $H_{i,0}$ if there exists one $|S_{i,j}(p)| \geq c_j(\alpha_i)$. This notation $c_j(\alpha_i)$ here is essentially c_j, but we want to emphasize that these thresholds will be later used to construct $(1-\alpha_i)$ level confidence sets. According to the duality between hypothesis test and confidence set, this $(1 - \alpha_i)$ level two-sided confidence set can be obtained by the acceptance region inversion of this group sequential test:

$$\mathcal{A}_i^{\text{gs}} = \bigcap_{j=1}^{G} \{|S_{i,j}(p)| < c_j(\alpha_i)\}\,.$$

Note that this confidence set is calculated at the end of the permutation tests. We hope that our confidence sets are all intervals. When this is true, followed by the monotonicity property (Tamhane and Liu, 2008; Gandy and Hahn, 2014) of BH procedure, an upper limit of interval being rejected by BH implies all elements from the interval are rejected by

BH; and a lower limit being accepted implies all elements from the interval are accepted. We can prove the following result related to the shape of $\mathcal{A}_i^{\text{gs}}$.

Theorem 4.1. For thresholds $c_1(\alpha_i), c_2(\alpha_i), \cdots, c_G(\alpha_i)$ that satisfy (4.3), an individual acceptance region inversion $\{|S_{i,j}(p)| < c_j(\alpha_i)\}$, $\{S_{i,j}(p) < c_j(\alpha_i)\}$, or $\{S_{i,j}(p) > -c_j(\alpha_i)\}$ must be an interval.

A corollary of this theorem is that their intersection $\mathcal{A}_i^{\text{gs}}$ must be an interval as well. It maintains the coverage probability over the nominal level for any stopping time that is arbitrarily chosen at j from $\{1, 2, \cdots, G\}$. In addition, with some additional derivation, it is easy to show that $\mathcal{A}_i^{\text{gs}} = \mathcal{A}_i^{\text{gs}+} \cap \mathcal{A}_i^{\text{gs}-}$ where the lower bound interval

$$\mathcal{A}_i^{\text{gs}+} = \bigcap_{j=1}^{G} \{S_{i,j}(p) < c_j(\alpha_i)\},$$

and the upper bound interval

$$\mathcal{A}_i^{\text{gs}-} = \bigcap_{j=1}^{G} \{S_{i,j}(p) > -c_j(\alpha_i)\}$$

are two $(1 - \alpha_i/2)$ level one-sided confidence intervals. It is worth mentioning that in our sequential algorithm, the one sided confidence interval estimated at the j-th group is written as $\cap_{l=1}^{j}\{S_{i,l} < c_l(\alpha_i)\}$ or $\cap_{l=1}^{j}\{S_{i,l} > -c_l(\alpha_i)\}$. For any stopping time, the coverage probability of these intervals should exceed the nominal level. Using upper bound interval as an example, this is because

$$\Pr\left(\bigcap_{j=1}^{G}\left\{p_i^* \in \bigcap_{l=1}^{j}\{S_{i,l} > -c_l(\alpha_i)\}\right\}\right) = \Pr\left(p_i^* \in \bigcap_{j=1}^{G}\{S_{i,j} > -c_j(\alpha_i)\}\right) \geq 1 - \alpha_i/2.$$

This result also works for lower bound interval. With some abuse of notation, $\mathcal{A}_i^{\text{gs}+}$ and $\mathcal{A}_i^{\text{gs}-}$ represent confidence intervals that we obtain throughout the process of permutation tests. From now on, they are not necessarily limited to the final intervals at the end of permutation tests. For both sets of thresholds from (4.4) and (4.5), it is easy to verify that $c_j(\alpha_i)$ monotonically increases when α_i decreases to 0. This implies that the interval $\mathcal{A}_i^{\text{gs}+}$ or $\mathcal{A}_i^{\text{gs}-}$ becomes wider with a smaller α_i. In the next section, we show that we need to

frequently update these intervals because of a step-wise decision process. At each testing group α_i is adjusted by the number of confirmed rejections and acceptances. Fortunately due to the closed-form thresholds in (4.4) and (4.5), it will only takes minimal time in computing these intervals.

4.2.3 Step-Wise Procedure to Control Family-Wise MCER

In addition to replacing Robbins-Lai interval with the ones developed by group sequential methods, the Bonferroni correction in Gandy and Hahn (2014, 2016) to achieve simultaneous coverage is actually not necessary. Alternatively we consider a step-wise procedure to construct simultaneous confidence intervals. Two-sided simultaneous intervals would be the best fit for our problem because MC error is a type of directional error. However, the general step-wise approach to develop two-sided simultaneous intervals has not been developed, or even possibly it does not exist. A famous counterexample was given by Shaffer (1980) to show that step-wise decision procedure could possibly fail to control a directional error rate. Concerning the theoretical difficulty, here we construct two sets of $(1-\alpha/2)$ level one-sided simultaneous intervals. The first set provides upper bound estimates for $\mathbf{p}^*$ and the second set provides lower bound estimates. Obviously the intersection of the two sets yields $(1-\alpha)$ level two-sided simultaneous intervals. This step-wise procedure is developed with the partition principle (Finner and Strassburger, 2002; Strassburger and Bretz, 2008).

We present our methods to compute one-sided (upper bound) simultaneous intervals $\times_{i=1}^{m}\mathcal{A}_i^{\text{gs}-} = \times_{i=1}^{m}(0, p_i^u)$, assuming that we are now at the end of the first group ($j=1$) of generating permutations. One-sided (lower bound) intervals $\times_{i=1}^{m}\mathcal{A}_i^{\text{gs}+}$ can be computed analogously. In the first step, we create $(1-\alpha/2)$ level one-sided simultaneous intervals $\times_{i=1}^{m}\mathcal{A}_i^{\text{gs}-}$ for all m tests with Bonferroni correction. Each $\mathcal{A}_i^{\text{gs}-}$ is computed at $(1-\alpha/(2m))$ level, using permutations from the first group. Then standard BH is applied on the upper limits of each $\mathcal{A}_1^{\text{gs}-}, \mathcal{A}_2^{\text{gs}-}, \cdots, \mathcal{A}_m^{\text{gs}-}$. We classify a test into the rejection category, if the corresponding upper limit is rejected by standard BH. Suppose that $m_{r,j}$ is the cumulative number of rejections at the end of the j-th group. If $m_{r,1}=0$ meaning that none test can be rejected, the process of step-wise interval estimation under $j=1$

Initialize $\widehat{D}_1, \widehat{D}_1, \cdots, \widehat{D}_m$ to be *undecided*.
Initialize $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ to be $(0, 1)$.
Initialize m_u to be m, and $m_{u'}$ to be $(m+1)$.

▷ Permutation begins

for *j in* 1 : G **do**
  while $m_u \neq m_{u'}$ **do**
    $m_u = \#\{i \in \{1, 2, \cdots, m\} \mid \widehat{D}_i = \textit{undecided}\}$.
    $m_r = \#\{i \in \{1, 2, \cdots, m\} \mid \widehat{D}_i = \textit{rejection}\}$.
    $m_a = \#\{i \in \{1, 2, \cdots, m\} \mid \widehat{D}_i = \textit{acceptance}\}$.
    ▷ The sum $m_u + m_r + m_a$ is always m
    $m_{u'} = m_u$
    for *i in* 1 : m **do**
      if $\widehat{D}_i =$ *undecided* **then**
        Find $\alpha^+ = \alpha/(m - m_a)$.
        Assign one-sided lower bound interval $\mathcal{A}_i^{\text{gs}+} = \bigcap_{l=1}^{j} \{S_{i,l}(p) < c_l(\alpha^+)\}$.
        Find $\alpha^- = \alpha/(m - m_r)$.
        Assign one-sided upper bound interval $\mathcal{A}_i^{\text{gs}-} = \bigcap_{l=1}^{j} \{S_{i,l}(p) > -c_l(\alpha^-)\}$.
        The two-sided group sequential interval $\mathcal{I}_i = \mathcal{A}_i^{\text{gs}+} \cap \mathcal{A}_i^{\text{gs}-}$.
      end
    end
    Apply standard BH on the upper limits of intervals $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ and decisions are $d_1^u, d_2^u, \cdots, d_m^u$.
    Apply standard BH on the lower limits of intervals $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ and decisions are $d_1^l, d_2^l, \cdots, d_m^l$.
    for *i in* 1 : m **do**
      if $d_i^u = d_i^l =$ *acceptance* **then**
        Update $\widehat{D}_i =$*acceptance*.
      end
      if $d_i^u = d_i^l =$ *rejection* **then**
        Update $\widehat{D}_i =$*rejection*.
      end
    end
  end
end

▷ Permutation ends

return $(\widehat{D}_1, \widehat{D}_1, \cdots, \widehat{D}_m)$.

Algorithm 4: Monte Carlo tests with group sequential approaches, which are our proposed methods for sequential resampling-based multiple testing.

will be terminated, and we will further estimate intervals under $j = 2$ after permutations of the second group arrive. If $m_{r,1} > 0$, the simultaneous intervals can be estimated with more than one step for $j = 1$. In the second step, the one-sided simultaneous confidence intervals will be updated by a $(1 - \alpha/(2m - 2m_{r,1}))$ level interval for each of those remaining $(m - m_{r,1})$ tests that are not rejected previously. Standard BH is applied again on the upper limits of the updated confidence intervals. For those tests that are rejected in the first step and not estimated again in the second step, we assign their upper limits based on the previous (first) step. If we observe at least one more rejection except those have been counted in $m_{r,1}$, we will classify them into rejection category, and proceed to a third step. The number of rejections $m_{r,1}$ will be updated at the same time. We keep with this rule. Excluding rejected tests in earlier steps, in a new step we only construct one-sided simultaneous confidence intervals on the remaining set of tests using the Bonferroni correction. A next step is performed if more rejections can be found. Otherwise we stop this step-wise process. At the time of exiting the step-wise process for the first $(j = 1)$ group, we move to the second $(j = 2)$ group with initializing $m_{r,2} = m_{r,1}$. The step-wise procedure for $j > 1$ is similar to the one for $j = 1$. We stop our algorithm after we reach the last group of permutations, or at an intermediate group if all hypotheses have been classified to rejection or acceptance category. More details can be found in Algorithm 4. The following theorem justifies the valid error rate control of our algorithm.

Theorem 4.2. Given that both $A_i^{\text{gs}+}$ and $A_i^{\text{gs}-}$ defined in Section 4.2.2 are $(1 - \alpha_i/2)$ level one-sided confidence intervals, the MCTGS (Algorithm 4) controls the family-wise MCER by α, and controls both family-wise MCER-I and family-wise MCER-II by $\alpha/2$.

4.3 Simulation Studies

4.3.1 Setup

We compared **MCTGS-flat**: our proposed MCTGS algorithm that implemented the flat thresholds defined in (4.4), **MCTGS-parabolic**: MCTGS algorithm that implemented the parabolic thresholds defined in (4.5), and **GH16**: the sequential resampling-based testing methods proposed by Gandy and Hahn (2016). The GH16 method can be adapted to our group sequential scheme. Suppose that G groups were considered in sequential resampling-based tests. With all methods we can decide *rejection*, *acceptance*, or *undecided* after collecting every (N_{max}/G) permutation samples on each hypothesis. We considered $m = 1000$ tests. We chose $N_{max} = 100,000$ which is usually a large number in a real setting. The number of groups, G could range from 20, 100, 500. The nominal level of family-wise MCER α was set to 20%; the nominal FDR q was set to 10%.

Except group design, we considered the similar simulation settings in Chapter 3. We chose the proportion of nulls $\pi_0 = 80\%$: $m\pi_0 = 800$ null p-values were sampled independently from U$[0, 1]$ distribution, and $m(1 - \pi_0) = 200$ non-null p-values were sampled independently from the Gaussian right-tailed probability model:

$$p_i^* = 1 - \Phi(Z_i), \qquad Z_i \sim \mathrm{N}(\beta, 1), \qquad i = m\pi_0 + 1,\ m\pi_0 + 2,\ \ldots,\ m.$$

The effect size β could be 1.5, 2 and 2.5. Given a fixed set of m, π_0 and β for sampling ideal p-values, we sampled 100 different realizations of $\mathbf{p}^*$. We showed in Chapter 3 that the number of rejections based on different realizations may vary from 0 to 160. Instead of generating $\mathbb{I}(T_{i,k} > T_i)$ in equation (1.3) through real permutations, we independently draw Bernoulli samples with success probability p_i^* for $i = 1, 2, \cdots, m$ and $k = 1, 2, \cdots, N_{max}$.

We evaluated two types of error rates empirically at the end of sequential resampling-based tests: the family-wise MCER and family-wise MCER-I. In addition to the error

rates, detection sensitivity defined in Chapter 3 and the percentage of decided tests

$$100\% - \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(\widehat{D}_i = \textit{undecided}),$$

were compared among our proposed methods and the competing methods. A higher percentage of decided tests can also suggest that a method is more powerful in terms of concluding rejection or acceptance decisions. Lastly, we compared the computational cost in permutation through the following metric called

$$\textit{percentage of permutations} = \frac{1}{m}\sum_{i=1}^{m}\frac{N_i}{N_{max}},$$

where N_i is the number of permutation samples that have been generated until the stopping time specifically for the i-th test. If we could assume that the time complexity of either MCTGS or GH16 is negligible compared to permutation, and treat time spent on permutation per sample per test as a constant, this metric would represent the ratio of actual running time between a sequential testing algorithm and the respective non-sequential testing algorithm that we have discussed in the last chapter. Each metric was evaluated by the average over 1,000 simulation replicates.

4.3.2 Simulation Results

We found that both empirical family-wise MCER and MCER-I were usually conservative compared to the nominal level. We observed none error in most cases.

Figure 4.2 displays that our method with either flat or parabolic thresholds is more powerful than the GH16 method in all scenarios. Adopting the flat thresholds achieves best detection sensitivity. In most cases, those tests which were only rejected by MCTGS-flat were undecided with other two approaches. We noticed that detection sensitivity under $\beta = 1.5$ is usually lower than $\beta = 2$ and 2.5, which is similar to the pattern of Figure 3.5. In addition, it seems that the choice of G does not have a strong impact on detection sensitivity. Possibly this is because the sequential confidence intervals are similar to the ones under a different size of groups.

From Figure 4.3, we observed that our MCTGS method significantly improves the

number of decided tests compared to the existing method. Using the flat thresholds yields more decided tests compared to the parabolic thresholds. Similar to the observation in Figure 4.2, the percentage of decided tests does not vary among different choices of G. Usually fewer than 2% of the tests cannot be decided to either rejection or acceptance region with MCTGS-flat, as p-values of these tests are very close to the BH cutoff.

Figure 4.4 shows the computational cost in generating permutation samples. MCTGS-parabolic usually takes the smallest number of permutations. Assuming that permutation contributes to the majority of the overall computational cost in sequential resampling-based tests, this further implies that MCTGS-parabolic is the most computationally efficient algorithm among our comparison, MCTGS-flat usually costs slightly more than the GH16 method. Possibly this type of sequential confidence interval is wider than its alternatives when we start running our algorithm. Recall that we mentioned the connection between thresholds (4.4) and group sequential literature. The group sequential test developed by O'Brien and Fleming (1979) is also known to perform similarly, usually being quite conservative at the early stages. However, as we accumulate more permutation samples, the cost in generating permutation samples should be similar between the two MCTGS methods.

4.4 Application to Prostate Cancer Data

We evaluated performances of the two versions of our MCTGS algorithm using the same prostrate cancer gene expression data from Chapter 3. We first picked parameter $N_{max} = 500,000$ and $1,000,000$ based on the power results of Section 3.4. In addition, the number of groups was set to 500. We also applied the GH16 method and the Sandve's MCFDR algorithm to this dataset. In terms of the Sandve's MCFDR algorithm, the only tuning parameter h was chosen to be 20 which is the default value suggested by Sandve et al. (2011). This implies that permutation sampling would not be stopped until we observe 20 permuted statistics that are larger than the observed test statistic. The nominal FDR for all methods was set to 10%. For MCTGS-flat, MCTGS-parabolic and the GH16 method, the nominal family-wise MCER was set to 20%. For comparison purposes, we obtained the testing results for 10 runs.

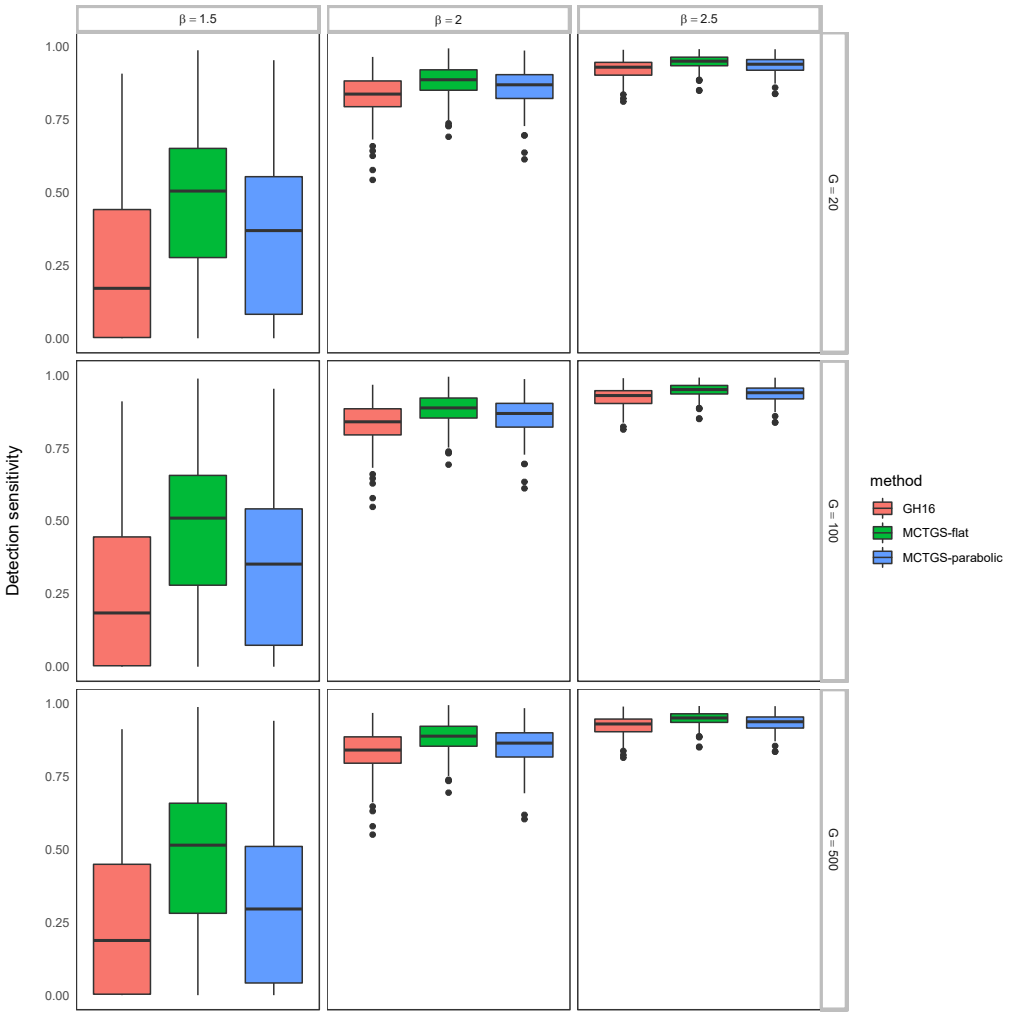

Figure 4.2: Detection sensitivity in all 100 different realizations using sequential resampling-based tests. The non-null ideal p-values were simulated from the Gaussian right-tailed model with $\beta = 1.5, 2.0$ and 2.5. The number of group was chosen from 20, 100 and 500.

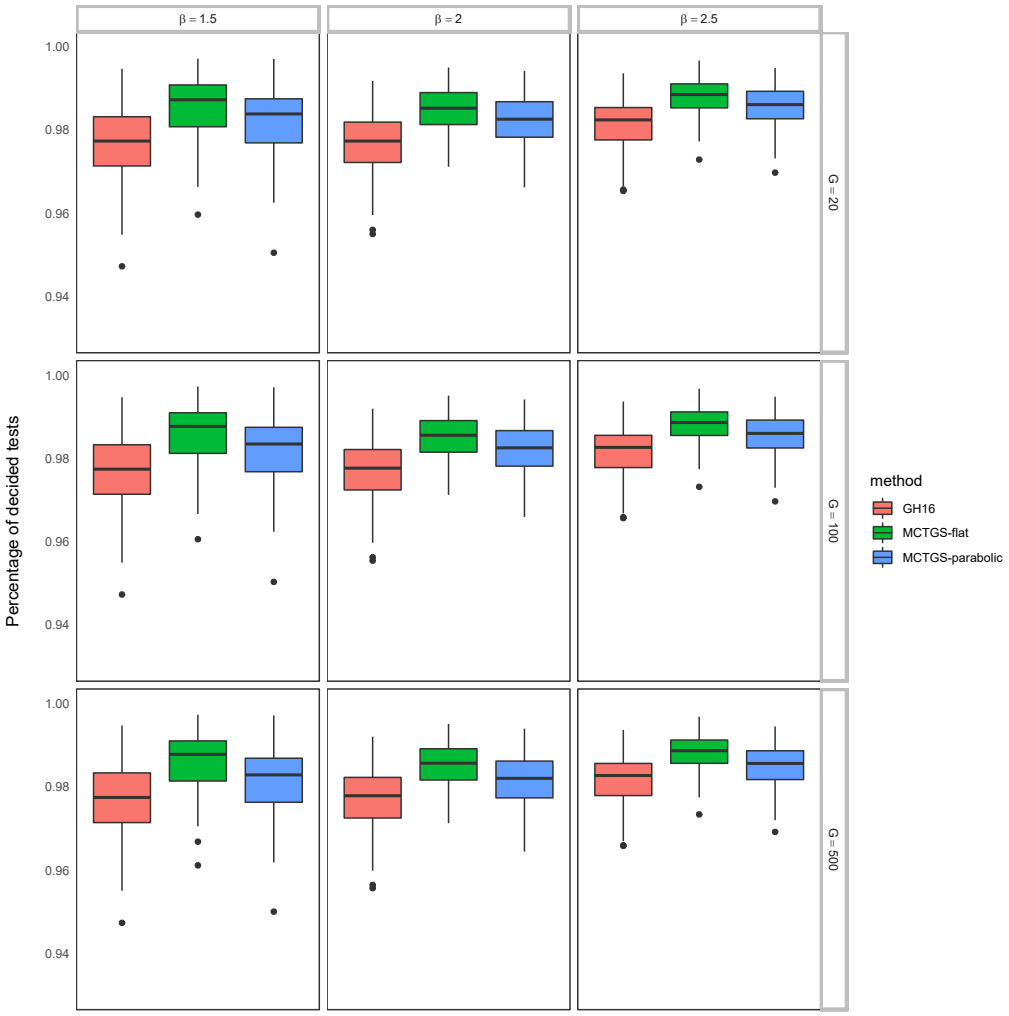

Figure 4.3: Percentage of decided tests in all 100 different realizations using sequential resampling-based tests. The non-null ideal p-values were simulated from the Gaussian right-tailed model with $\beta = 1.5, 2.0$ and 2.5. The number of group was chosen from 20, 100 and 500.

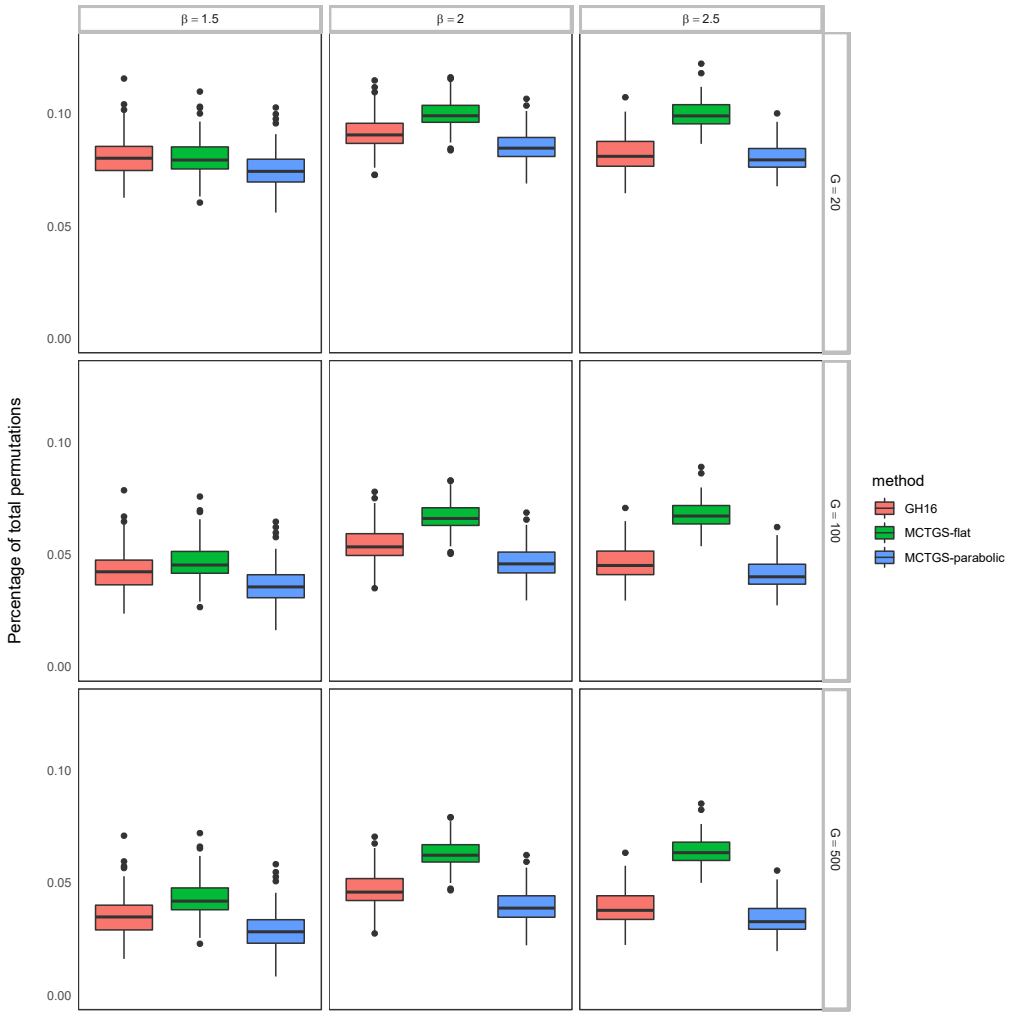

Figure 4.4: Percentage of actual permutation samples generated in sequential resampling-based tests compared to N_{max} in all 100 different realizations. The non-null ideal p-values were simulated from the Gaussian right-tailed model with $\beta = 1.5, 2.0$ and 2.5. The number of group was chosen from 20, 100 and 500.

Under the scenarios where $N_{max} = 500,000$ or $N_{max} = 1,000,000$, we observed 17 tests that flipped their decisions (from *rejection* to *acceptance* or vice versa) at least once among the 10 runs with Sandve's MCFDR, which indicates the occurrence of MC errors. None flipped decision was observed with the remaining three methods. In Table 4.1 and 4.2, among the list of significant DE genes and not significant genes by the two versions of MCTGS, the chance that there exists a rejection or acceptance decision which would not be reproduced if we run the permutation for an infinite number is less than 20% in each experiment.

In each run, both MCTGS-flat and MCTGS-parabolic methods yield more rejection and acceptance decisions than the GH16 method. The number of undecided tests with MCTGS-flat is here at most half of the number with GH16 method. In addition, the MCTGS-flat method yields a fewer number of undecided tests than MCTGS-parabolic.

Finally, MCFDR always demonstrates the lowest computational cost in permutation sampling. However, as a trade-off, there is no promise on the precision of their decisions. Among the two MCTGS methods and the GH16 method, the computational costs in terms of permutation sampling are similar.

Table 4.1: The number of detected DE genes, non-DE genes, undecided genes, and the computational costs in permutation when $N_{max} = 500,000$ and $\alpha = 20\%$.

	Run	1	2	3	4	5	6	7	8	9	10
MCTGS-flat											
	rejected	56	54	56	51	52	53	56	54	51	53
	accepted	5965	5969	5964	5967	5969	5970	5968	5965	5969	5968
	undecided	12	9	13	15	12	10	9	14	13	12
	perm. cost	1.6%	1.5%	1.5%	1.5%	1.5%	1.6%	1.6%	1.5%	1.5%	1.5%
MCTGS-para.											
	rejected	48	51	54	51	50	50	51	50	50	51
	accepted	5962	5967	5964	5966	5966	5966	5964	5965	5966	5963
	undecided	23	15	15	16	17	17	18	18	17	19
	perm. cost	1.0%	0.9%	0.9%	0.9%	1.0%	1.0%	1.0%	0.9%	0.9%	0.9%
GH16											
	rejected	47	47	48	49	50	47	49	48	50	49
	accepted	5960	5959	5959	5961	5961	5961	5959	5960	5962	5960
	undecided	26	27	26	23	22	25	25	25	21	24
	permu. cost	1.2%	1.2%	1.1%	1.1%	1.2%	1.2%	1.2%	1.2%	1.2%	1.2%
Sandve's MCFDR											
	rejected	61	63	62	56	61	61	63	60	58	64
	accepted	5972	5970	5971	5977	5972	5972	5970	5973	5975	5969
	undecided	0	0	0	0	0	0	0	0	0	0
	perm. cost	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%

Table 4.2: The number of detected DE genes, non-DE genes, undecided genes, and the computational costs in permutation when $N_{max} = 1,000,000$ and $\alpha = 20\%$.

	Run	1	2	3	4	5	6	7	8	9	10
MCTGS-flat											
	rejected	57	57	57	57	57	58	57	58	57	57
	accepted	5967	5970	5971	5971	5971	5971	5971	5971	5969	5970
	undecided	9	6	5	5	5	4	5	4	7	6
	perm. cost	1.2%	1.1%	1.1%	1.1%	1.1%	1.1%	1.1%	1.1%	1.1%	1.1%
MCTGS-para.											
	rejected	55	56	57	57	57	56	57	57	57	56
	accepted	5967	5970	5966	5971	5966	5971	5968	5968	5969	5968
	undecided	11	7	10	5	10	6	8	8	7	9
	perm. cost	0.7%	0.6%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%
GH16											
	rejected	53	55	57	52	55	54	52	52	51	54
	accepted	5962	5963	5964	5964	5965	5963	5964	5964	5964	5963
	undecided	18	15	12	17	13	16	17	17	18	16
	permu. cost	0.9%	0.8%	0.8%	0.8%	0.8%	0.9%	0.9%	0.8%	0.8%	0.8%
Sandve's MCFDR											
	rejected	62	63	62	56	61	61	63	60	58	64
	accepted	5971	5970	5971	5977	5972	5972	5970	5973	5975	5969
	undecided	0	0	0	0	0	0	0	0	0	0
	perm. cost	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%

4.5 Discussion

In this work, we derived a class of sequential resampling-based multiple testing algorithms. For sequential multiple testing, we are able to stop resampling in an early stage as long as the hypothesis has been assigned a rejection or acceptance decision. Meanwhile, our sequential procedure provides theoretical guarantee on reproducibility on those tests that are either rejected or accepted. It can be shown to control the same error rate, family-wise MCER, defined by Gandy and Hahn (2014, 2016). Most importantly, our methods achieve higher power at the end of permutation tests compared to the existing methods. The improved power is related to two factors: more powerful sequential confidence intervals which are based on group sequential testing, as well as the step-wise decision procedure. In terms of the concept of reproducibility, some researchers might expect that with an algorithm which enjoys a perfect property of reproducibility, two separate runs of the algorithm should generate the exactly same results with a high probability, even on those undecided tests. Unfortunately, none of the existing methods (including our methods) can provide such a strong promise with theoretical justification. However, it should be an interesting question for future research.

Our methods can incorporate many other types of group sequential approaches. In

addition to the standard designs of Pocock (1977) and O'Brien and Fleming (1979), more types of designs could be generated by the so called α-spending functions (Demets and Lan, 1994). The thresholds c_j can be solved by numerical integration approaches (Genz and Bretz, 2009), although it would take much more computational cost compared to our current algorithms that only consider closed-form thresholds. It would be an interesting question to find the optimal set of thresholds which can achieve best detection sensitivity under general situations. Classic group sequential methods require specifying the number of groups and the decision threshold for each group at the design stage (the period before running an actual test). In our case, however, this requirement implies that we cannot generate extra permutation samples at the end of tests even when some hypotheses still remain undecided. In literature, solutions to such a problem are usually called adaptive group sequential approaches (Bauer and Kohne, 1994; Lehmacher and Wassmer, 1999; Müller and Schäfer, 2001). The sequential confidence interval based on these more flexible alternatives was considered by Mehta et al. (2007), which we believe can be incorporated into our methods. We are extending our framework in these promising directions.

Chapter 5

Summary and Future Directions

In this dissertation I proposed three novel statistical methods for multiple-hypotheses testing. The main contribution of this dissertation is three-folded. First, we translated complex scientific problems into well-defined multiple testing problems and clarified an error rate we hoped to control in each problem. Second, we theoretically showed that our proposed methods control the error rates below the desired level. Third, our proposed methods always demonstrate higher statistical power than both existing methods and naïve solutions to the problem.

In Chapter 2, we discussed the problem of testing associated taxa in microbiome data. We defined a novel error rate called false assignment rate that is an intermediate surrogate between FDR under the global null and FDR under the conjunction null. Both versions of FDR have their own limitations and thus are not the ideal candidate for this problem. We proposed a bottom-up decision rule and proved that under certain conditions, this rule provides the exact control of false assignment rate. We used simulation studies and real data application to show that our methods are more capable of identifying driver nodes of a microbial community than those existing methods which test global null hypothesis at each high level node. It also provides a framework for testing tree-structured multiple hypotheses, especially when global null does not define the right scientific question. We mentioned that the independence assumption is required for this framework. First, we need this assumption to obtain the exact p-values of Stouffer's Z tests at upper levels. Liu and Xie (2019) proposed a Cauchy combination test for aggregating individual p-values. The Cauchy combination test was shown to be robust under arbitrary dependence structures. In preliminary studies, we observe that replacing Stouffer's test with Cauchy combination test achieves much better error rate control under our existing bottom-up framework, even when there are strong pairwise correlations among leaf nodes. Second, the step-down multiple testing procedures usually requires independence to control FDR rigorously. Some simulation results in Gavrilov et al. (2009) suggested that empirical FDR would be inflated under strong pairwise correlation between test statistics. Therefore, developing a bottom-up test that can be adapted to general dependency structure is a meaningful but challenging question. A weaker question is to control the error rate asymptotically. It seems that bootstrap and subsampling methods (Romano et al., 2008) could be used to account for dependency structure and control FDR in the asymptotic

sense, although large computational complexity is still a concern.

In Chapter 3 and Chapter 4, we focused on the problem of controlling Monte Carlo errors for multiple resampling-based testing. We showed that existing methods based on the Bonferroni's simultaneous confidence intervals are conservative and lack of power. We proposed two improved alternatives in both sequential and non-sequential scenarios. However, we observed in simulation studies that the empirical distribution of MCER is highly skewed. For most $\mathbf{p}^*$, the error rates are nearly 0, which means the methods are still conservative. Under very sporadic cases, the error rates could be close to the nominal level. Similar phenomena occur in many other multiple testing problems (such as testing directions of multiple treatment effects) where the null hypotheses are not point-wise but composite nulls. To avoid this issue, we could possibly consider the average MCER or MCER-I as the error rate that we would like to control. In many scientific problems (Pounds and Morris, 2003; Chung et al., 2014), we can assume the ideal p-values of individual tests follow certain parametric distribution. When the parametric assumption can be justified, the family-wise MCER-I could be replaced by error rate $\int V^{\text{MC-I}} d\pi(p^*)$ which is an integral of type-I MC error over $\pi(p^*)$, the sampling distribution of ideal p-values $\mathbf{p}^*$. This can be viewed as a starting point towards an empirical Bayes solution. The next step would be finding a systematic way in estimating the distribution function of $\mathbf{p}^*$. In Efron (2016), this type of estimation was called Bayesian deconvolution.

Testing multiple omnibus hypotheses can be useful in many biomedical applications. For example, when testing association in microbiome data, it is known that one type of association test is usually more powerful in some scenarios whereas it could be less powerful in other scenarios. The omnibus association test which combines multiple types of association tests into one, significantly enhances power (Koh et al., 2017; Hu and Satten, 2017). However, the most powerful procedure is usually unknown, especially under the scenarios we do not necessarily emphasize type-I error control on individual tests but instead hope to control FDR on the list of discoveries. A multiple testing procedure that directly tests multivariate statistics can be more powerful than a univariate procedure (e.g., BH procedure) that is applied on the p-values of omnibus tests. Some related discussions could be found in Du et al. (2014); Zhao (2015); Alishahi et al. (2016).

Appendix A

Appendix for Chapter 2

A.1 Proof of Theorem 2.1

The following lemma is essentially the summation by parts formula and will be used in the proof of Theorem 2.1.

Lemma A1. Suppose $\{a_1, \ldots, a_n\}$ and $\{b_1, \ldots, b_n\}$ are two sets of real numbers. Then,

$$\sum_{k=1}^{n} a_k b_k = \sum_{k=1}^{n} (a_k - a_{k-1}) B_k,$$

where $B_k = \sum_{i=k}^{n} b_i$ and $a_0 = 0$.

Proof of Theorem 2.1. This proof is adapted from Gavrilov et al. (2009) with modifications for our multi-level, bottom-up procedure. Let R_l and V_l denote the number of detection and false detection, respectively, under the modified null hypothesis at level l. Then the false assignment proportion (FAP) is written as

$$\text{FAP} = \frac{\sum_{l=1}^{L} V_l}{(\sum_{l=1}^{L} R_l) \bigvee 1} \leq \sum_{l=1}^{L} \frac{V_l}{(\sum_{l'=1}^{l} R_{l'}) \bigvee 1} = \sum_{l=1}^{L} \frac{V_l}{(D_{l-1} + R_l) \bigvee 1} = \sum_{l=1}^{L} \frac{V_l}{D_{l-1} + (R_l \bigvee 1)}.$$

The key step to prove Theorem 2.1 is to show that for every level l

$$\mathbb{E}\left[\frac{V_l}{D_{l-1} + (R_l \bigvee 1)} \middle| \mathcal{G}_{l-1} \right] \leq q_l, \tag{A1}$$

where $\mathcal{G}_{l-1}$ represents detection events below level l. The inequality (A1) does not guarantee the control of FAR at each level l because of the cumulative effect of D_{l-1}, which establishes a dependence between the nodes detected at different levels. However, it leads to the control of overall FAR at q:

$$\text{FAR} = \mathbb{E}(\text{FAP}) \leq \sum_{l=1}^{L} \mathbb{E}\left\{\mathbb{E}\left[\frac{V_l}{D_{l-1} + (R_l \bigvee 1)}\middle|\mathcal{G}_{l-1}\right]\right\} \leq \sum_{l=1}^{L} q_l = q.$$

To prove (A1), we omit the level index l for simplicity of exposition. Thus we redefine the lth-level p-values to be $p_1, \ldots, p_{n^*}$, ordered p-values $p_{(1)} \leq \cdots \leq p_{(n^*)}$, and thresholds defined in (2.1) $\alpha_1 \leq \cdots \leq \alpha_{n^*}$. We use D_{-1} for D_{l-1}. We also omit $\mathcal{G}_{l-1}$ by acknowledging that the ensuing arguments are always conditional on the detection events below level l. Further, we denote the set $\left\{p_{(1)} \leq \alpha_1, \ldots, p_{(k)} \leq \alpha_k\right\}$ by $\mathcal{A}_k$ $(k = 1, \ldots, n^*)$, which represents the case that the first k ordered p-values are each below the first k thresholds.

To start, we use the definitions of V_l and R_l and rewrite the left hand side of (A1) to be

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \mathbb{E}\left[\frac{\mathbb{I}\left(\mathcal{A}_k, p_{(k+1)} > \alpha_{k+1}, p_j \leq \alpha_k\right)}{D_{-1} + k}\right], \tag{A2}$$

where $\mathcal{H}_0$ denotes the set of modified null hypotheses at level l. Then, we replace the expectation by double expectations that first conditions on p_j and apply Lemma A1 with $a_k = \mathbb{I}\left(p_j \leq \alpha_k\right)/(D_{-1} + k)$, $b_k = \Pr\left(\mathcal{A}_k, p_{(k+1)} > \alpha_{k+1}\middle|p_j\right)$, and $n = n^*$; note that $b_{n^*} = \Pr\left(\mathcal{A}_{n^*}\middle|p_j\right)$. Thus, (A2) becomes

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \mathbb{E}\left\{\Pr\left(\mathcal{A}_k|p_j\right) \times \left[\frac{\mathbb{I}\left(p_j \leq \alpha_k\right)}{D_{-1} + k} - \frac{\mathbb{I}\left(p_j \leq \alpha_{k-1}\right)}{D_{-1} + k - 1}\right]\right\},$$

which can be reorganized as

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \left[\frac{\Pr\left(\mathcal{A}_k, \alpha_{k-1} < p_j \leq \alpha_k\right)}{D_{-1} + k} - \frac{\Pr\left(\mathcal{A}_k, p_j \leq \alpha_{k-1}\right)}{(D_{-1} + k)(D_{-1} + k - 1)}\right]. \tag{A3}$$

Let $p_{(1)}^{(-j)} \leq \cdots \leq p_{(n^*-1)}^{(-j)}$ be the ordered p-values after excluding p_j. We denote the set $\left\{p_{(1)}^{(-j)} \leq \alpha_1, \ldots, p_{(k-1)}^{(-j)} \leq \alpha_{k-1}\right\}$ by $\mathcal{B}_{k-1}^{(-j)}$, which represents the case that the first $(k-1)$ ordered p-values after excluding p_j are each below the first $(k-1)$ thresholds. We note

two facts that relate $\mathcal{A}_k$ and $\mathcal{B}_{k-1}^{(-j)}$. First, $\mathcal{A}_k$ and $\{\alpha_{k-1} < p_j\}$ together imply that p_j cannot be among the first $(k-1)$ smallest p-values and thus the first $(k-1)$ ordered p-values before and after excluding p_j remain the same set. In addition, for any $j \in \{(k), \ldots, (n^*)\}$, $\{p_j \leq \alpha_k\}$ implies $\left\{p_{(k)} \leq \alpha_k\right\}$. Thus we have $\Pr\left(\mathcal{A}_k, \alpha_{k-1} < p_j \leq \alpha_k\right) = \Pr\left(\mathcal{B}_{k-1}^{(-j)}, \alpha_{k-1} < p_j \leq \alpha_k\right)$. Second, $\mathcal{B}_{k-1}^{(-j)}$ is a subset of $\mathcal{A}_k$ when $p_j \leq \alpha_{k-1}$, which yields $\Pr\left(\mathcal{A}_k, p_j \leq \alpha_{k-1}\right) \geq \Pr\left(\mathcal{B}_{k-1}^{(-j)}, p_j \leq \alpha_{k-1}\right)$. By the two facts, we see that expression (A3) is less than

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \left[\frac{\Pr\left(\mathcal{B}_{k-1}^{(-j)}, \alpha_{k-1} < p_j \leq \alpha_k\right)}{D_{-1}+k} - \frac{\Pr\left(\mathcal{B}_{k-1}^{(-j)}, p_j \leq \alpha_{k-1}\right)}{(D_{-1}+k)(D_{-1}+k-1)} \right],$$

which, by the independence assumption between p_j and all other p-values given $\mathcal{G}_{l-1}$, becomes

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \Pr\left(\mathcal{B}_{k-1}^{(-j)}\right) \times \left[\frac{\Pr\left(p_j \leq \alpha_k\right)}{D_{-1}+k} - \frac{\Pr\left(p_j \leq \alpha_{k-1}\right)}{D_{-1}+k-1} \right].$$

Applying Lemma A1 with $a_k = \Pr\left(p_j \leq \alpha_k\right)/(D_{-1}+k)$ and $b_k = \Pr\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k\right)$, the foregoing expression reduces to

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \Pr\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k\right) \times \frac{\Pr\left(p_j \leq \alpha_k\right)}{D_{-1}+k}. \tag{A4}$$

Now we find (A4) to be an upper bound for (A2).

By the uniform distribution of p_j for $j \in \mathcal{H}_0$ and the definition of the thresholds $\{\alpha_k\}$, we have $\Pr\left(j \leq \alpha_k\right)/(D_{-1}+k) = \alpha_k/(D_{-1}+k) \leq q_l(1-\alpha_k)/(n^*+1-k) = q_l \Pr\left(p_j > \alpha_k\right)/(n^*+1-k)$. In addition, replacing $\sum_{j \in \mathcal{H}_0}$ by $\sum_{j=1}^{n^*}$ in (A4) yields an upper bound for (A4):

$$q_l \sum_{k=1}^{n^*} (n^*+1-k)^{-1} \sum_{j=1}^{n^*} \Pr\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k, p_j > \alpha_k\right).$$

We see that $\mathcal{B}_{k-1}^{(-j)}$ and $\{p_j > \alpha_k\}$ imply that p_j is not among the top $(k-1)$ smallest p-values. For either $j = (k)$ or $j \in \{(k+1), \ldots, (n^*)\}$, we infer from $p_{(k)}^{(-j)} > \alpha_k$ and $p_j > \alpha_k$ that $p_{(k)} > \alpha_k$. Therefore, $\Pr\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k, p_j > \alpha_k\right) = \Pr\left(\mathcal{A}_{k-1}, p_{(k)} > \alpha_k\right)$ for $j \in \{(k), \ldots, (n^*)\}$ and 0 for $j \in \{(1), \ldots, (k-1)\}$, which simplifies the above expression

to

$$q_l \sum_{k=1}^{n^*} \Pr\left(\mathcal{A}_{k-1}, p_{(k)} > \alpha_k\right) = q_l \left[1 - \Pr\left(\mathcal{A}_{n^*}\right)\right] \leq q_l.$$

We complete the proof of (A1). □

A.2 Proof of Theorem 2.2

We first find the mathematical form for weight ω_j (after omiting the level index l from the original notation $\omega_{l,j}$) for node j at level l. Recall that the p-value at node j is denoted by p_j. Let $T_1, T_2, \ldots, T_t$ represent all subtrees that contain node j. For example, the node $N_{1,1}$ in Figure 2.1(a) is contained in subtrees rooted at $N_{2,1}$, $N_{3,1}$ and $N_{4,1}$, which are denoted by T_1, T_2 and T_3, respectively. Let $p_{T_s}^{(-j)}$ be the set of p-values of all level-l nodes that are contained in subtree T_s excluding node j. First, each weight ω_j always starts with 1 indicating the node itself. Then for each subtree T_s, if p_j is the (only) maximum p-value, i.e., $p_j > p_{T_s}^{(-j)}$, rejecting node j will entail the root node of that subtree to also be rejected and thus add 1 indicating that root node to ω_j. Therefore, we can write ω_j as summation of a series of indicator functions:

$$\omega_j = 1 + \sum_{s=1}^{t} \mathbb{I}\left(p_j > p_{T_s}^{(-j)}\right). \tag{A5}$$

Note that ω_j is dependent on p-values at level l and thus considered to be random (even given the detection events below level l).

Lemma A2 states a property of ω_j, which will be useful in the proof of Theorem 2.2.

Lemma A2. Suppose node j has weight ω_j as defined in (A5). Assume that node j is under the null hypothesis and so its p-value p_j follows the uniform distribution. Also assume that p_j is independent of all other p-values at level l. Let $\mathcal{B}^{(-j)}$ denote the case that all other p-values excluding p_j belong to a Borel set. For any $\alpha \in (0, 1)$, we have

$$\mathbb{E}\left[\omega_j \mathbb{I}\left(\mathcal{B}^{(-j)}, p_j \leq \alpha\right)\right] \leq \frac{\alpha}{1-\alpha} \mathbb{E}\left[\omega_j \mathbb{I}\left(\mathcal{B}^{(-j)}, p_j > \alpha\right)\right].$$

Proof of Lemma A2. Assume that there is no tie in p-values. By the independence assumption of p_j and the other p-values and the uniform distribution of p_j, we have

$\mathbb{E}\left[\mathbb{I}\left(\mathcal{B}^{(-j)}, p_j \le \alpha\right)\right]/\alpha = \Pr\left(\mathcal{B}^{(-j)}\right) = \mathbb{E}\left[\mathbb{I}\left(\mathcal{B}^{(-j)}, p_j > \alpha\right)\right]/(1-\alpha)$. Due to the linearity of ω_j, it then suffices to show for any subtree T_s that

$$\frac{\mathbb{E}\left[\mathbb{I}\left(\mathcal{B}^{(-j)}, p_j > p_{T_s}^{(-j)}, p_j \le \alpha\right)\right]}{\Pr\left(p_j \le \alpha\right)} \le \frac{\mathbb{E}\left[\mathbb{I}\left(\mathcal{B}^{(-j)}, p_j > p_{T_s}^{(-j)}, p_j > \alpha\right)\right]}{\Pr\left(p_j > \alpha\right)}. \tag{A6}$$

By the mean value theorem, there exist p_j^* and p_j^{**}, where $0 < p_j^* < \alpha < p_j^{**} < 1$, such that the left hand side of (A6) becomes

$$\int_0^\alpha \Pr\left(\mathcal{B}^{(-j)}, p_j > p_{T_s}^{(-j)} \mid p_j\right) dF(p_j) \Big/ \int_0^\alpha 1 dF(p_j) = \Pr\left(\mathcal{B}^{(-j)}, p_j^* > p_{T_s}^{(-j)}\right)$$

and the right hand side becomes

$$\int_\alpha^1 \Pr\left(\mathcal{B}^{(-j)}, p_j > p_{T_s}^{(-j)} \mid p_j\right) dF(p_j) \Big/ \int_\alpha^1 1 dF(p_j) = \Pr\left(\mathcal{B}^{(-j)}, p_j^{**} > p_{T_s}^{(-j)}\right).$$

The fact that $p_j^* \le p_j^{**}$ gives us (A6). □

Proof of Theorem 2.2. We introduce some new notation related to the weights $\omega_1 \ldots, \omega_{n^*}$ for level-l nodes. We denote the relative ordering of p-values $p_1, \ldots, p_{n^*}$ by $\mathcal{O}$. The weights defined in (A5) are thus uniquely determined given the detection events at lower levels $\mathcal{G}_{l-1}$ as well as the ordering $\mathcal{O}$. Let $\omega_{[j]}$ be the weight corresponding to the j-th smallest p-value $p_{(j)}$ and $C_k = \sum_{j=1}^k \omega_{[j]}$ for $k = 1, \ldots, n^*$. Thus, C_k is also deterministic given $\{\mathcal{G}_{l-1}, \mathcal{O}\}$. Let $\omega_{(1)} \le \omega_{(2)} \le \cdots \le \omega_{(n^*)}$ denote the sorted weights by their own values; note that $\omega_{(j)}$ is often different from $\omega_{[j]}$. As illustrated in Section 2.1.4, $\omega_{(j)}$ is deterministic given $\mathcal{G}_{l-1}$ only under Condition (C1), regardless of the ordering of p-values. Denote $c_k = \sum_{j=1}^k \omega_{(j)}$ and $\bar{c}_k = \sum_{j=k}^{n^*} \omega_{(j)}$. Hence, $c_k \le C_k$, $\bar{c}_k \ge C_{n^*-k+1}$, and the threshold α_k satisfies

$$\frac{\alpha_k}{1-\alpha_k} \le \frac{D_{-1} + \sum_{j=1}^k \omega_{(j)}}{\sum_{j=k}^{n^*} \omega_{(j)}} q_l = \frac{D_{-1} + c_k}{\bar{c}_k} q_l. \tag{A7}$$

Like the proof of Theorem 2.1, the key step is to show that for every level l

$$\mathbb{E}\left[\frac{V_l}{D_{l-1} + (R_l \bigvee 1)} \middle| \mathcal{G}_{l-1}\right] \le q_l. \tag{A8}$$

Again, we omit the index l and the condition $\mathcal{G}_{l-1}$ and denote the set $\left\{p_{(1)} \leq \alpha_1, \ldots, p_{(k)} \leq \alpha_k\right\}$ by $\mathcal{A}_k$. The left hand side of (A8) can be rewritten as

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \mathbb{E}\left[\frac{\omega_j \mathbb{I}\left(\mathcal{A}_k, p_{(k+1)} > \alpha_{k+1}, p_j \leq \alpha_k\right)}{D_{-1} + C_k}\right],$$

where C_k by definition counts the number of all rejections that are entailed by rejecting the k-th smallest p-value. Replacing the expectation by double expectations that first conditions on $\mathcal{O}$ yields

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \mathbb{E}_{\mathcal{O}}\left\{\omega_j \mathbb{E}\left[\frac{\mathbb{I}\left(\mathcal{A}_k, p_{(k+1)} > \alpha_{k+1}, p_j \leq \alpha_k\right)}{D_{-1} + C_k}\middle| \mathcal{O}\right]\right\}, \tag{A9}$$

where C_k becomes a constant given $\mathcal{O}$.

Once we condition on the order $\mathcal{O}$ and the weights $\{w_j\}$ become constants, similar arguments used in steps between expressions (A2) and (A4) in proof of Theorem 2.1 can also be used to obtain an upper bound for (A9):

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \mathbb{E}_{\mathcal{O}}\left\{\omega_j \mathbb{E}\left[\frac{\mathbb{I}\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k, p_j \leq \alpha_k\right)}{D_{-1} + C_k}\middle| \mathcal{O}\right]\right\}. \tag{A10}$$

Next, we combine the double expectations in (A10) into one and use $c_k \leq C_k$ to find that (A10) is less than

$$\sum_{j \in \mathcal{H}_0} \sum_{k=1}^{n^*} \left(D_{-1} + c_k\right)^{-1} \mathbb{E}\left[\omega_j \mathbb{I}\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k, p_j \leq \alpha_k\right)\right]. \tag{A11}$$

Now the weight ω_j is considered random again. According to Lemma A2 with $\mathcal{B}^{(-j)} = \left\{\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k\right\}$, $\alpha = \alpha_k$, and a null p-value p_j, we obtain

$$\mathbb{E}\left[\omega_j \mathbb{I}\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k, p_j \leq \alpha_k\right)\right] \leq \frac{\alpha_k}{1 - \alpha_k} \mathbb{E}\left[\omega_j \mathbb{I}\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k, p_j > \alpha_k\right)\right].$$

Using (A7) and replacing $\sum_{j \in \mathcal{H}_0}$ by $\sum_{j=1}^{n^*}$, we see (A11) is less than

$$q_l \sum_{k=1}^{n^*} \left(\overline{c}_k\right)^{-1} \mathbb{E}\left[\sum_{j=1}^{n^*} \omega_j \mathbb{I}\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k, p_j > \alpha_k\right)\right].$$

By the same arguments as in proof of Theorem 2.1, $\mathbb{I}\left(\mathcal{B}_{k-1}^{(-j)}, p_{(k)}^{(-j)} > \alpha_k, p_j > \alpha_k\right) = \mathbb{I}\left(\mathcal{A}_{k-1}, p_{(k)} > \alpha_k\right)$ for $j \in \{(k), \ldots, (n^*)\}$ and 0 for $j \in \{(1), \ldots, (k-1)\}$. Therefore, the above expression simplifies to

$$q_l \sum_{k=1}^{n^*} (\overline{c}_k)^{-1} \mathbb{E}\left[\overline{s}_k \mathbb{I}\left(\mathcal{A}_{k-1}, p_{(k)} > \alpha_k\right)\right] \leq q_l \sum_{k=1}^{n^*} \Pr\left(\mathcal{A}_{k-1}, p_{(k)} > \alpha_k\right) = q_l \left[1 - \Pr\left(\mathcal{A}_{n^*}\right)\right] \leq q_l.$$

We complete the proof of (A8). □

A.3 Least Favorable Weights

We obtain $\widetilde{\boldsymbol{\omega}} = (\widetilde{\omega}_{l,(1)}, \cdots, \widetilde{\omega}_{l,(n_l^*)})$ in a recursive manner. Recall that a weight $\omega_{l,j}$ counts the number of nodes that are simultaneously rejected if node j is rejected, which includes node j and some of its ancestors. Let $\widetilde{\boldsymbol{\omega}}^{(l+h)}$ $(h = 0, 1, 2, \ldots)$ denote the least favorable set of weights against any possible set of weights $\boldsymbol{\omega}^{(l+h)}$ when nodes between level l and level $(l+h)$ (including levels l and $l+h$) are counted. Trivially, $\widetilde{\boldsymbol{\omega}}^{(l)} = \boldsymbol{\omega}^{(l)} = (1, \ldots, 1)$ with 1 for every node when only nodes at level l are counted, and this serves as the starting point of the recursive algorithm. At the root node level, denoted by $(l + h^*)$, $\widetilde{\boldsymbol{\omega}}^{(l+h^*)}$ is the least favorable set of weights $\widetilde{\boldsymbol{\omega}}$ that we wish to obtain, and is the end point of the algorithm. We find that $\widetilde{\boldsymbol{\omega}}^{(l+h)}$ can be derived from $\widetilde{\boldsymbol{\omega}}^{(l+h-1)}$ by traversing over every node at level $l + h$, locating the subset of elements in $\widetilde{\boldsymbol{\omega}}^{(l+h-1)}$ that correspond to the level-l descendants of that node, and adding 1 to the largest element of that subset; if there exist multiple largest elements, randomly pick one and add 1. For example, to calculate least favorable weights for the bottom level of the tree in Figure 2.1(b), we obtain $\widetilde{\boldsymbol{\omega}}^{(l)} = (1, 1, 1, 1, 1, 1)$, $\widetilde{\boldsymbol{\omega}}^{(l+1)} = (1, 2, 1, 1, 1, 2)$, and $\widetilde{\boldsymbol{\omega}}^{(l+2)} = (1, 2, 1, 1, 1, 3)$, and ultimately the sorted version $\widetilde{\boldsymbol{\omega}} = (1, 1, 1, 1, 2, 3)$. Note that we ordered the individual weights in intermediate $\widetilde{\boldsymbol{\omega}}^{(l+h)}$ by the physical position of the bottom-level nodes in the displayed tree, and only sort the weights in the last step. This procedure guarantees that every $\widetilde{\boldsymbol{\omega}}^{(l+h)}$ $(h = 0, 1, \ldots, h^*)$ is the least favorable against any arbitrary $\boldsymbol{\omega}^{(l+h)}$ in which the count 1 is added to one element other than the largest one for at least one subset. It also ensures the uniqueness of sorted $\widetilde{\boldsymbol{\omega}}$, which leads to a unique set of thresholds. Finally, the least favorable weights $\widetilde{\boldsymbol{\omega}}$ corresponds to the ordering of p-values that is equal to the ordering of depths at level-l nodes, e.g., the node with the largest p-value is the node

with the largest depth.

A.4 Proof of Theorem 2.3

Proof of Theorem 2.3. Let $\widetilde{\omega}_{(1)} \leq \widetilde{\omega}_{(2)} \leq \cdots \leq \widetilde{\omega}_{(n^*)}$ denote sorted least favorable weights after omitting the level index l. Denote $\widetilde{c}_k = \sum_{j=1}^{k} \widetilde{\omega}_{(j)}$ and $\overline{\widetilde{c}}_k = \sum_{j=k}^{n^*} \widetilde{\omega}_{(j)}$. The same arguments in the proof of Theorem 2.2 can be used except that we use $\widetilde{c}_k$ in place of c_k, $\overline{\widetilde{c}}_k$ in place of $\overline{c}_k$, and the thresholds (2.5) in place of thresholds (2.4). □

A.5 Proof of Theorem 2.4

Proof of Theorem 4. The OTU-level testing in the two-stage procedure is exactly the same as the OTU-level testing in the one-stage procedure, so we immediately have $\mathrm{FAR}_{\mathrm{otu}} \leq q_1$. Then we rewrite the FAP among all taxa (inner) levels as

$$\mathrm{FAP}_{\mathrm{taxa}} = \frac{\sum_{l=2}^{L} V_l}{(\sum_{l=2}^{L} R_l) \bigvee 1} \leq \sum_{l=2}^{L} \frac{V_l}{(\sum_{l'=2}^{l} R_{l'}) \bigvee 1} = \sum_{l=2}^{L} \frac{V_l}{(D_{l-1}^{\dagger} + R_l) \bigvee 1} = \sum_{l=2}^{L} \frac{V_l}{D_{l-1}^{\dagger} + (R_l \bigvee 1)}.$$

We note that $D_{l-1}^{\dagger}$ is deterministic conditioning on the detection events at lower levels, denoted by $\mathcal{G}_{l-1}^{\dagger}$, which excludes the OTU level. Then we follow the same steps in the proofs of Theorems 2.2 and 2.3, replacing D_{l-1} by $D_{l-1}^{\dagger}$ and $\mathcal{G}_{l-1}$ by $\mathcal{G}_{l-1}^{\dagger}$ to obtain $\mathbb{E}\left[V_l / \left\{D_{l-1}^{\dagger} + (R_l \bigvee 1)\right\} \middle| \mathcal{G}_{l-1}^{\dagger}\right] \leq q_l$ for $l = 2, \ldots, L$. Finally,

$$\mathrm{FAR}_{\mathrm{taxa}} = \mathbb{E}\left(\mathrm{FAP}_{\mathrm{taxa}}\right) \leq \sum_{l=2}^{L} \mathbb{E}\left\{\mathbb{E}\left[\frac{V_l}{D_{l-1}^{\dagger} + (R_l \bigvee 1)} \middle| \mathcal{G}_{l-1}^{\dagger}\right]\right\} \leq \sum_{l=2}^{L} q_l = q_{-1},$$

which implies that FAR among all taxa levels are controlled by q_{-1}. Indeed, this is the same as applying the one-stage testing at FAR q_{-1} to the subtree after removing the whole OTU level and the higher-level taxa that are detected because all of their corresponding OTUs are detected. □

Supplementary Materials for Simulations and Real Data

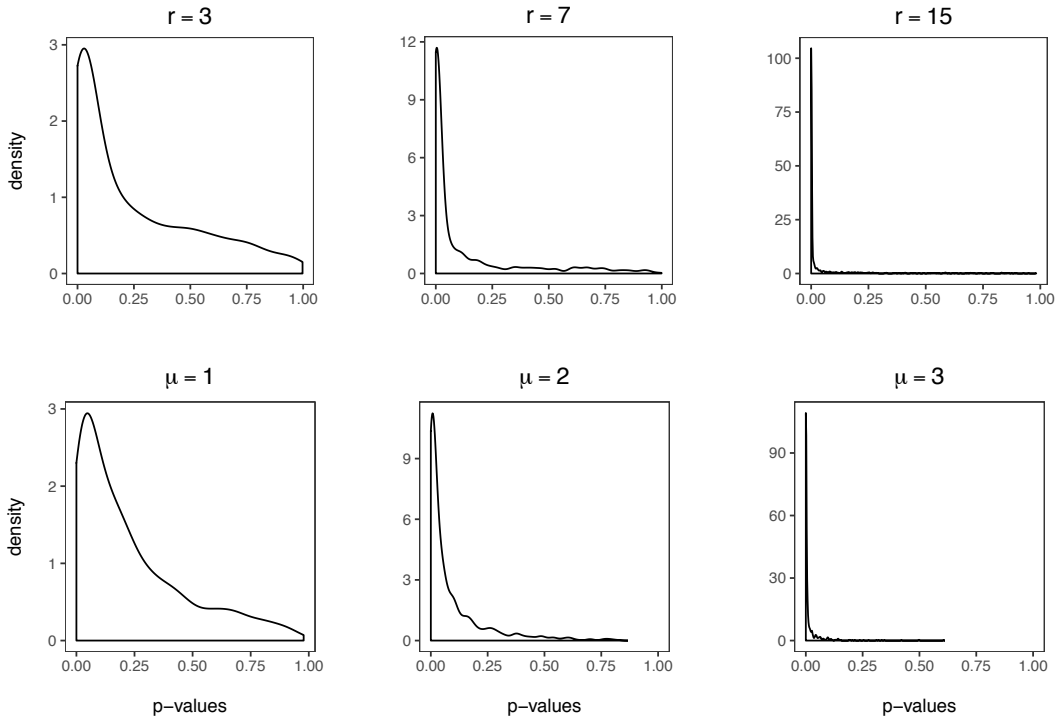

Figure A.1: Density functions of p-values generated from the Beta distribution Beta$(1/r, 1)$ (upper panel) and the Gaussian-tailed model (lower panel). The two plots in each column have comparable "height" at around zero but the upper ones always have heavier right tails.

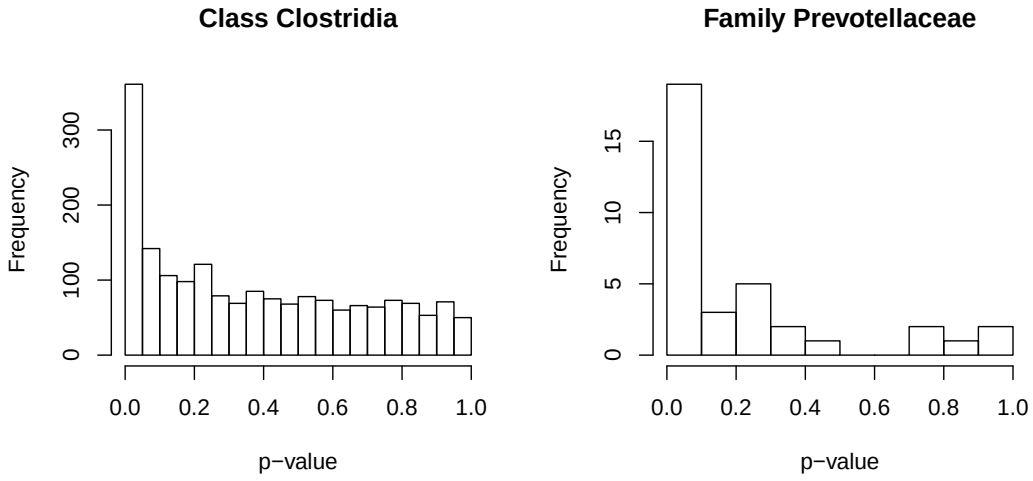

Figure A.2: Empirical distributions of p-values for OTUs in class *Clostridia* and family *Prevotellaceae* in the IBD data. These two taxa were detected to be driver taxa by the weighted bottom-up test and contain the most OTUs.

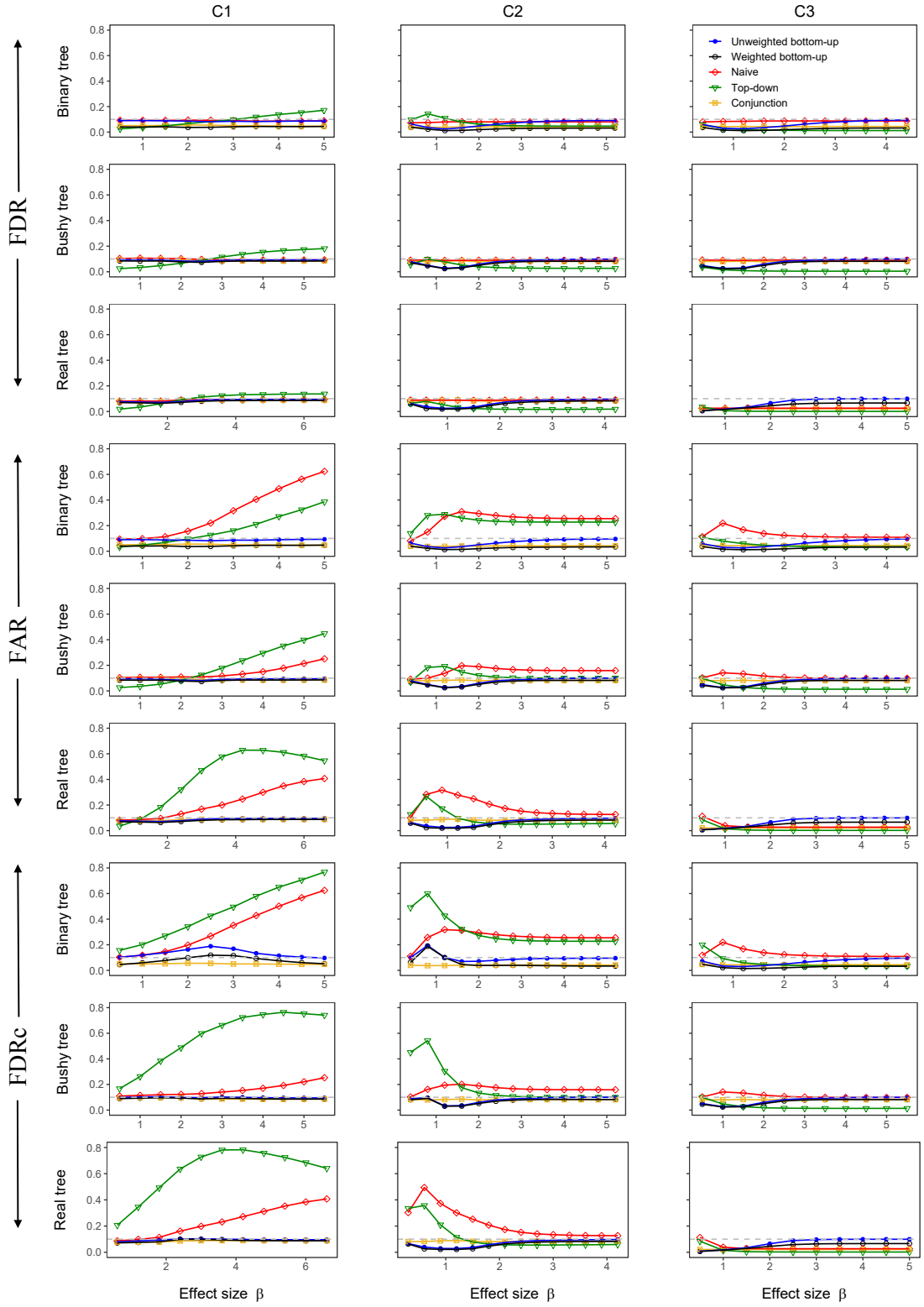

Figure A.3: Error rates for testing all nodes in the tree. The non-null p-values at leaf nodes were simulated from the Gaussian-tailed model.

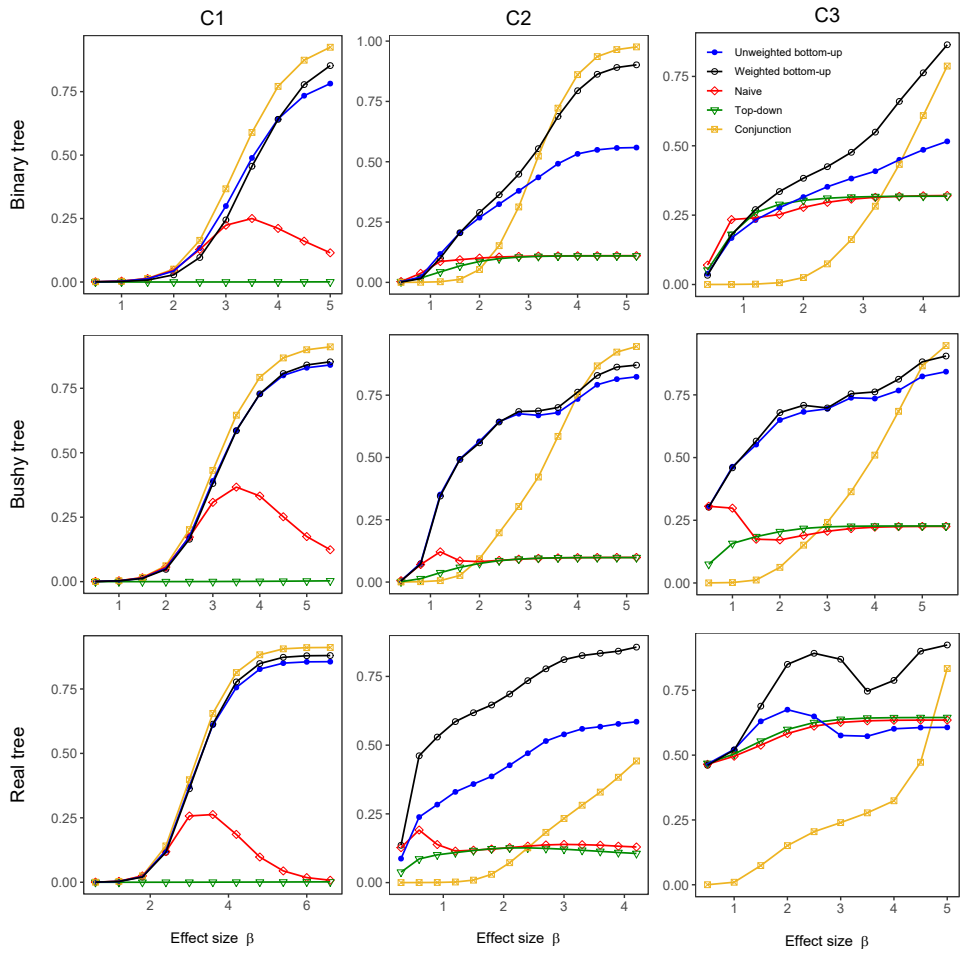

Figure A.4: Accuracy (weighted Jaccard similarity) for detecting all associated nodes (including the driver nodes and all of their descendants at all levels). The non-null p-values at leaf nodes were simulated from the Gaussian-tailed model.

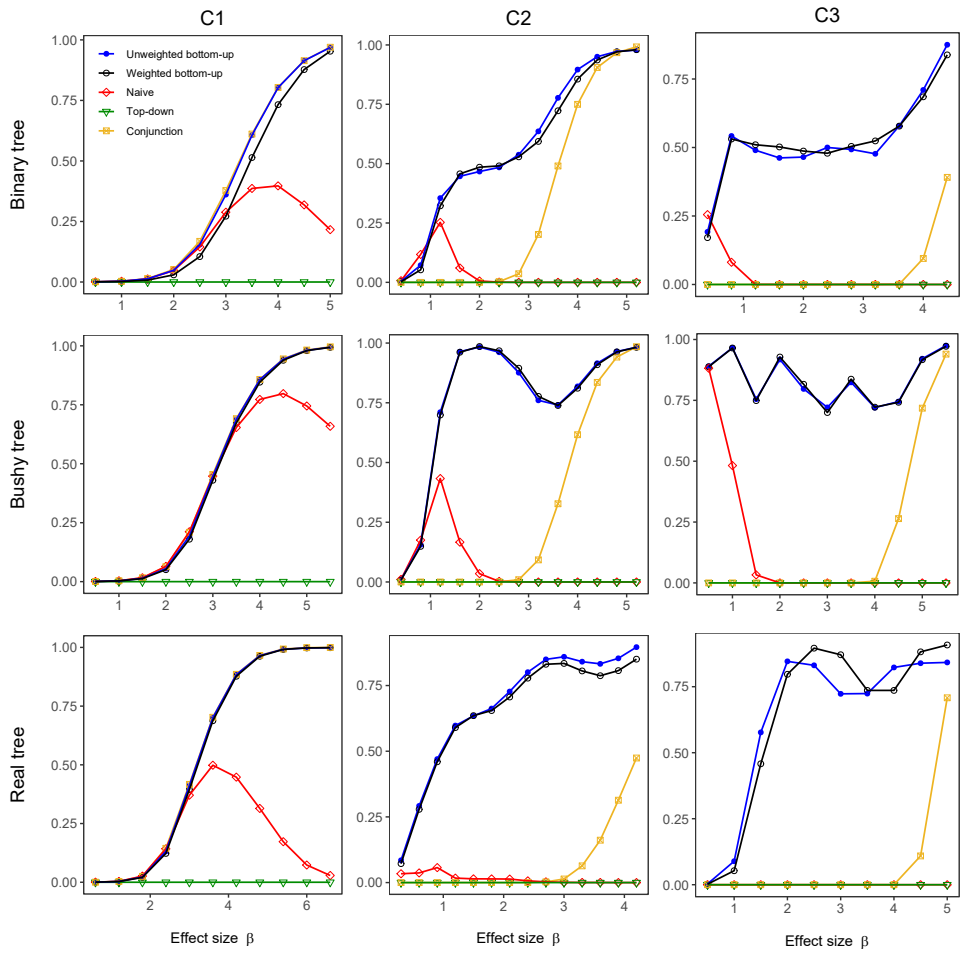

Figure A.5: Percentage of driver nodes that were pinpointed. The non-null p-values at leaf nodes were simulated from the Gaussian-tailed model.

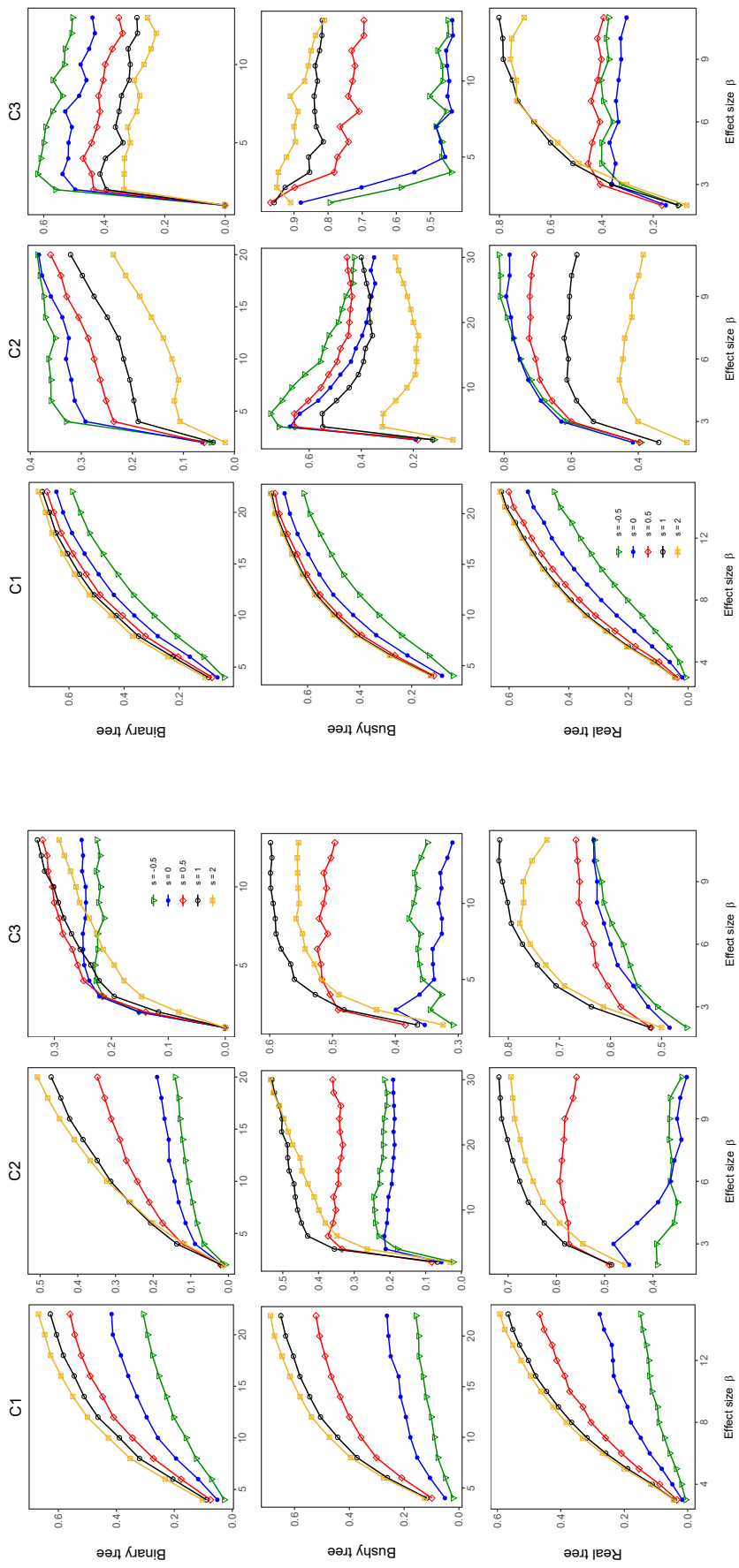

Figure A.6: Accuracy for detecting all associated nodes (left panel), and percentage of driver nodes that were pinpointed (right panel) by our weighted bottom-up test with different partitioning schemes that partition the total error rate q into $q_1, \ldots, q_L$. We considered partitioning schemes that each q_l is proportional to n_l^s, where $s \in \{-1/2, 0, 1/2, 1, 2\}$. Thus, $s = 1$ corresponds to the default partition in our bottom-up test. Compared to the default partition, $s = 2$ gives higher q_l to lower levels and $s = -1/2$ gives higher q_l to higher levels. We evaluated these schemes using all scenarios that we have considered in simulation; in particular, the non-null p-values at leaf nodes were simulated from the Beta distribution. We can see from the left panel that the default partition usually achieves the top-two best accuracy; and from the right panel that in terms of detecting driver nodes, the optimal partition varies in different scenarios and the default partition achieves the most robust performance over all scenarios.

Table A.1: Taxa detected by the weighted bottom-up test to be differential abundant between the UC and control groups

Taxon	p-value
p_Actinobacteria; *c_Coriobacteriia*$^{\#}$	1.39×10^{-4}
p_Actinobacteria; *c_Coriobacteriia*$^{\#}$; *o_Coriobacteriales*	1.39×10^{-4}
p_Actinobacteria; *c_Coriobacteriia*$^{\#}$; *o_Coriobacteriales*; *f_Coriobacteriaceae*	1.39×10^{-4}
p_Actinobacteria; *c_Coriobacteriia*$^{\#}$; *o_Coriobacteriales*; *f_Coriobacteriaceae*; *g_Slackia*	3.42×10^{-4}
p_Bacteroidetes; *c_Bacteroidia*; *o_Bacteroidales*; *f_Bacteroidaceae*; *g_Bacteroides*; *s_ovatus*$^{\#}$	3.09×10^{-5}
p_Bacteroidetes; *c_Bacteroidia*; *o_Bacteroidales*; *f_[Paraprevotellaceae]*; *g_[Prevotella]*$^{\#}$	3.44×10^{-4}
p_Bacteroidetes; *c_Bacteroidia*; *o_Bacteroidales*; *f_Prevotellaceae*$^{\#}$	1.64×10^{-3}
p_Bacteroidetes; *c_Bacteroidia*; *o_Bacteroidales*; *f_Prevotellaceae*$^{\#}$; *g_Prevotella*	1.64×10^{-3}
p_Bacteroidetes; *c_Bacteroidia*; *o_Bacteroidales*; *f_Prevotellaceae*$^{\#}$; *g_Prevotella*; *s_copri*	2.82×10^{-10}
p_Bacteroidetes; *c_Bacteroidia*; *o_Bacteroidales*; *f_S24-7*$^{\#}$	2.08×10^{-3}
p_Cyanobacteria; *c_Chloroplast*$^{\#}$	1.73×10^{-4}
p_Cyanobacteria; *c_Chloroplast*$^{\#}$; *o_Streptophyta*	1.73×10^{-4}
p_Firmicutes; *c_Clostridia*$^{\#}$	6.35×10^{-5}
p_Firmicutes; *c_Clostridia*$^{\#}$; *o_Clostridiales*	$< 1 \times 10^{-16}$
p_Firmicutes; *c_Clostridia*$^{\#}$; *o_Clostridiales*; *f_Clostridiaceae*; *g_Clostridium*	5.13×10^{-4}
p_Firmicutes; *c_Clostridia*$^{\#}$; *o_Clostridiales*; *f_Lachnospiraceae*	1.29×10^{-13}
p_Firmicutes; *c_Clostridia*$^{\#}$; *o_Clostridiales*; *f_Lachnospiraceae*; *g_Coprococcus*	3.20×10^{-7}
p_Firmicutes; *c_Clostridia*$^{\#}$; *o_Clostridiales*; *f_Ruminococcaceae*	$< 1 \times 10^{-16}$
p_Firmicutes; *c_Clostridia*$^{\#}$; *o_Clostridiales*; *f_Ruminococcaceae*; *g_Faecalibacterium*; *s_prausnitzii*	9.50×10^{-5}
p_Firmicutes; *c_Clostridia*$^{\#}$; *o_Clostridiales*; *f_Ruminococcaceae*; *g_Ruminococcus*	2.82×10^{-3}
p_Firmicutes; *c_Clostridia*$^{\#}$; *o_Clostridiales*; *f_Veillonellaceae*; *g_Veillonella*; *s_parvula*	1.00×10^{-4}
p_Firmicutes; *c_Erysipelotrichi*$^{\#}$	6.01×10^{-4}
p_Firmicutes; *c_Erysipelotrichi*$^{\#}$; *o_Erysipelotrichales*	6.01×10^{-4}
p_Firmicutes; *c_Erysipelotrichi*$^{\#}$; *o_Erysipelotrichales*; *f_Erysipelotrichaceae*	6.01×10^{-4}
p_Firmicutes; *c_Erysipelotrichi*$^{\#}$; *o_Erysipelotrichales*; *f_Erysipelotrichaceae*; *g_cc_115*	2.80×10^{-3}
p_Proteobacteria; *c_Gammaproteobacteria*; *o_Enterobacteriales*; *f_Enterobacteriaceae*; *g_Morganella*$^{\#}$	3.09×10^{-3}
p_Proteobacteria; *c_Gammaproteobacteria*; *o_Enterobacteriales*; *f_Enterobacteriaceae*; *g_Enterobacter*; *s_radicincitans*$^{\#}$	2.61×10^{-4}
p_Tenericutes; *c_RF3*$^{\#}$	4.25×10^{-3}
p_Verrucomicrobia$^{\#}$	1.25×10^{-3}
p_Verrucomicrobia$^{\#}$; *c_Verrucomicrobiae*	1.25×10^{-3}
p_Verrucomicrobia$^{\#}$; *c_Verrucomicrobiae*; *o_Verrucomicrobiales*	1.25×10^{-3}
p_Verrucomicrobia$^{\#}$; *c_Verrucomicrobiae*; *o_Verrucomicrobiales*; *f_Verrucomicrobiaceae*	1.25×10^{-3}
p_Verrucomicrobia$^{\#}$; *c_Verrucomicrobiae*; *o_Verrucomicrobiales*; *f_Verrucomicrobiaceae*; *g_Akkermansia*	1.25×10^{-3}
p_Verrucomicrobia$^{\#}$; *c_Verrucomicrobiae*; *o_Verrucomicrobiales*; *f_Verrucomicrobiaceae*; *g_Akkermansia*; *s_muciniphila*	1.25×10^{-3}

NOTE: The detected driver taxa are marked with "#". Kingdom *Bacteria* is omitted from the taxon names.

Appendix B

Appendix for Chapter 3

B.1 The Comparison between Robbins-Lai Interval and Wilson Interval

We chose ideal p-values p^* ranging from 0 to 0.2. Large p-values were not shown here because usually they are not recognized by BH as significant ones. We considered $n = 100,000$ permutations. We implemented $(1-\alpha)$ level two-sided Wilson confidence interval and $(1-\alpha)$ level two-sided Robbins-Lai interval at the time of ending all permutations. For each p^*, we calculated the expected upper limits of both confidence intervals with 1,000 simulations, and then found the ratio of two expected values.

In Figure B.1, the ratio of two expected upper limits is always above 1, which means that Wilson confidence interval always has a shorter expected upper limit than the Robbins-Lai interval. When p^* is approaching 0, the difference of length between two intervals are more significant. For large p^*, the upper limits of two intervals seem similar.

B.2 Splitting the Error Rate in the Two-Step Approach

We repeated the simulation studies with different ways of allocating error rates in the two steps. In this example, the total error rates is 10%. We considered 5 different plans to assign error rates separately to each step. First, the '1+9' plan means that the nominal error rates assigned to the first step = 1% and the nominal error rates assigned to the second step = 9%. Followed by this rule, we can similarly define other four plans, namely '3+7', '5+5', '7+3' and '9+1' plans. Obviously, our default choice of error rates split is the same as this '5+5' plan.

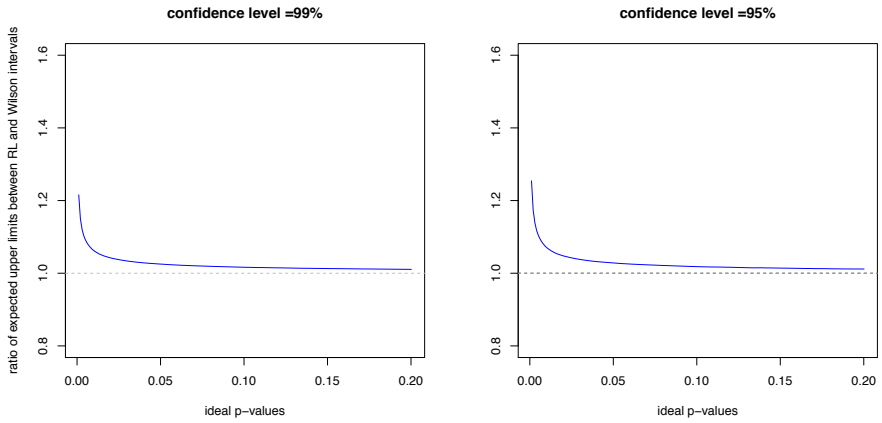

Figure B.1: The ratio of expected upper limits of two-sided Robbins-Lai and Wilson interval for $p^* \in (0, 0.2)$. We considered 99% level two-sided confidence interval in the left panel and 95% interval in the right panel.

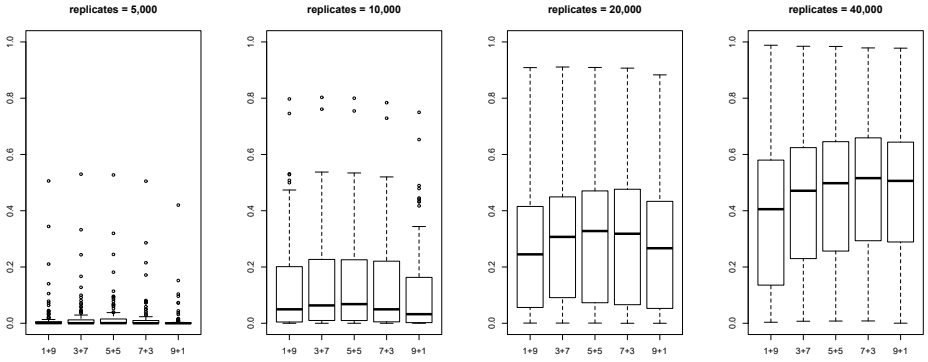

Figure B.2: The detection sensitivity under different splitting of the error rate between the two steps. The non-null ideal p-values were simulated from the Gaussian right-tailed model with $m = 1000, \beta = 1.5$.

Simulation results can be found in Figure B.2 and B.3. Apparently either '1+9' or '9+1' plan usually shows the lowest power. We see that there is more variation in power with all different allocation plans, if the number of permutations is relatively small. Under that scenario, our default choice, '5+5' shows the best power. However, when the number of replicates increases, all five allocation plans give quite similar power. As mentioned in previous sections, we would recommend to use the '5+5' plan in most cases.

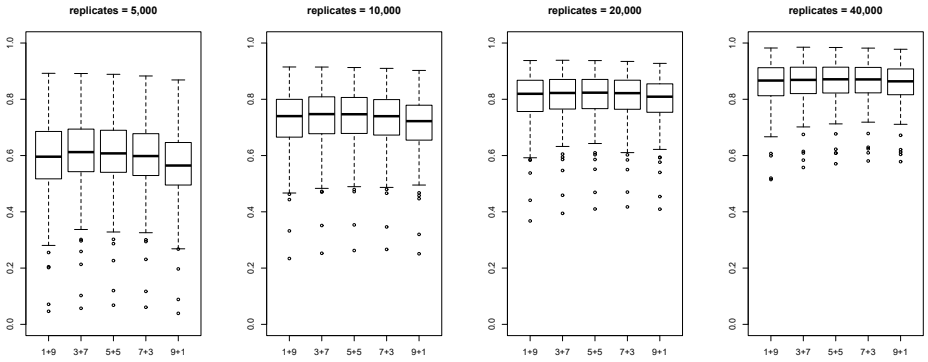

Figure B.3: The detection sensitivity under different splitting of the error rate between the two steps. The non-null ideal p-values were simulated from the Gaussian right-tailed model with $m = 1000, \beta = 2.0$.

B.3 Supplementary Materials for Simulations

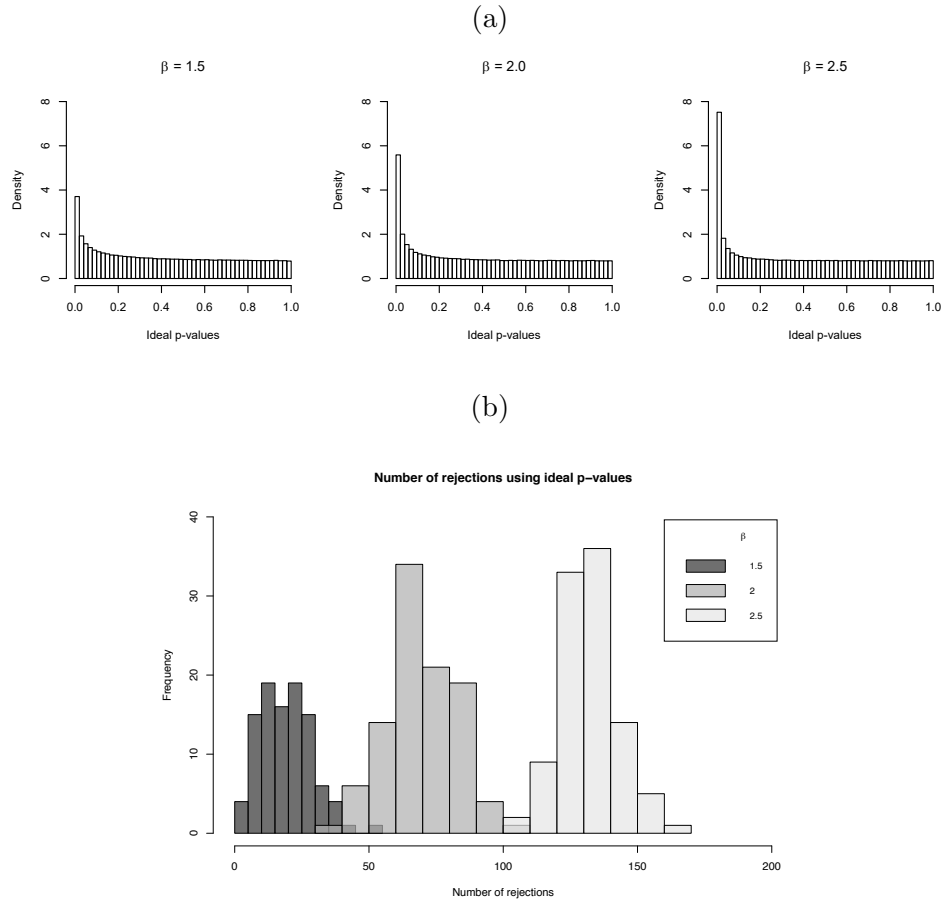

Figure B.4: (a) The theoretical distribution of the ideal p-values under the Gaussian right-tailed model with $\pi_0 = 0.8$. From left to right, effect size $\beta = 1.5, 2.0, 2.5$. (b) The empirical distribution of number of rejected hypotheses in the first set of simulations, given 1,000 ideal p-values simulated from the Gaussian right-tailed model.

B.4 The Extension of Two-Step Algorithm That Controls Family-Wise Type-II MCER

Algorithm 4 The two-step approach for controlling type-II MCER

Input: the matrix $\boldsymbol{I}$, nominal FDR rate q, nominal family-wise MCER-I α

Step 1: Constructing the $(1-\beta)$ level confidence interval $[0, \tau_u]$ for τ^*.

(1.1) For each $b \in (1, 2, \ldots, B)$, generate a bootstrap sample $(X_1^{(b)}, X_2^{(b)}, \ldots, X_m^{(b)})$ based on $\boldsymbol{I}^{(b)}$.

(1.2) Find $c^{(b)}$ to be the smallest c that satisfies $\min_{t\in[0,q]} \left\{F_c^{(b)}(t) - \widehat{F}(t)\right\} \geq 0$, where the shrinkage estimator $p_{c,i} = \frac{X_i - c\sqrt{X_i+1}}{n - c\sqrt{X_i+1}}$ and $F_c^{(b)}$ is now the empirical distribution function of $\frac{X_i^{(b)} - c\sqrt{X_i^{(b)}+1}}{n - c\sqrt{X_i^{(b)}+1}}$.

(1.3) Find c_u as the $(1-\beta)$ quantile of $(c^{(1)}, c^{(2)}, \ldots, c^{(B)})$.

(1.4) Find τ_u by applying the BH procedure on shrinkage estimators indexed by c_u.

Step 2: Testing multiple hypotheses $\widetilde{H}_{i,0} : p_i^* \leq \tau_u$ versus $\widetilde{H}_{i,1} : p_i^* > \tau_u$ given τ_u.

(2.1) If $\frac{X_{(i)} - n\tau_u}{\sqrt{\tau_u(1-\tau_u)}} \geq -\sqrt{2n\log(\log(n))}$, assign $\widetilde{p}_{(i)} = \tau_u$. Otherwise $\widetilde{p}_{(i)} = \frac{X_{(i)}}{n}$. For $i = m, (m-1), \ldots, 1$, if $\sum_{i'=1}^{i} \Pr\left\{\text{Bin}[n, \widetilde{p}_{(i')}] \geq X_{(i)}\right\} \leq \alpha - \beta$, reject $\widetilde{H}_{(i),0}$, update $i = i-1$, and rerun (2.1); otherwise, accept all remaining $\widetilde{H}_{i,0}$ and stop.

(2.2) If $i < 1$ then stop.

(2.3) Whenever $\widetilde{H}_{i,0}$ is rejected, $H_{i,0}$ is accepted.

Output: a list of accepted hypotheses by standard BH.

We evaluated its performance in the context of type-II MC error control using an almost identical simulation setup in Section 3.3, except that the number of permutation samples was here chosen from 2500, 5000 and 10000. In addition, we considered two competing methods. First, we slightly modified the GH-fixed method to fit this problem. It only differs from the version of Section 3.3.1 in that here we need the simultaneous lower bound intervals instead of upper bound intervals. We apply the standard BH on these lower limits, obtain the acceptance set $\mathcal{A}_0$ and decide the hypotheses in $\mathcal{A}_0$ to

be accepted. Second, we considered a variant of the proposed method in Chapter 4 which adopts the step-wise decision procedure. However, we no longer need sequential confidence intervals because in this problem the number of permutation samples is fixed. We call this method MCTGS-fixed. To be more specific, in the first step of this method, we construct one-sided simultaneous lower bound confidence intervals on all hypotheses with the Bonferroni correction. By applying BH, we accept those hypotheses whose corresponding upper limits of intervals are accepted by BH. In the k-th step, we only construct one-sided simultaneous lower bound intervals on the set of hypotheses that have not been accepted before. Using the updated confidence intervals, we similarly conclude hypotheses to be accepted. If there exists a hypothesis that is accepted in the k-th step but not in the $(k-1)$-th step, we continue to the $(k+1)$-th step; otherwise we terminate the entire step-wise procedure. In the final output, a hypothesis is claimed to be accepted if it is accepted in at least one step of running this algorithm. For both competing methods, all simultaneous confidence intervals are Wilson intervals. Similar to the notion of detection sensitivity in Section 3.1.1, we considered non-detection sensitivity

$$\mathrm{NS} = \frac{\sum_{i=1}^{m} \mathbb{I}(\widehat{D}_i = \mathit{acceptance}, D_i^* = \mathit{acceptance})}{1 \vee \sum_{i=1}^{m} \mathbb{I}(D_i^* = \mathit{acceptance})}.$$

It represents the average proportion of "ground truth" accepted hypotheses that are found by each method.

The results of non-detection sensitivity with all three methods are shown in Figure B.5. The two-step method achieved the highest non-detection sensitivity in all 9 scenarios, while GH-fixed was always the least powerful method among the three. For the two-step and MCTGS-fixed methods, we also assessed the enhancement of non-detection sensitivity compared to GH-fixed (i.e., $(\mathrm{NS}/\mathrm{NS}_0 - 1) \times 100\%$ where NS_0 is the non-detection sensitivity with GH-fixed). Results are shown in Figure B.6.

It is worthy noting that we can combine both Algorithm 3 and Algorithm 4 to control the family-wise MCER by 2α. Then a test will be assigned a *rejection* decision only if the output of Algorithm 3 indicates rejection; and will be assigned an *acceptance* decision only if the output of Algorithm 4 indicates acceptance. All remaining tests (if any) are labeled with *undecided*.

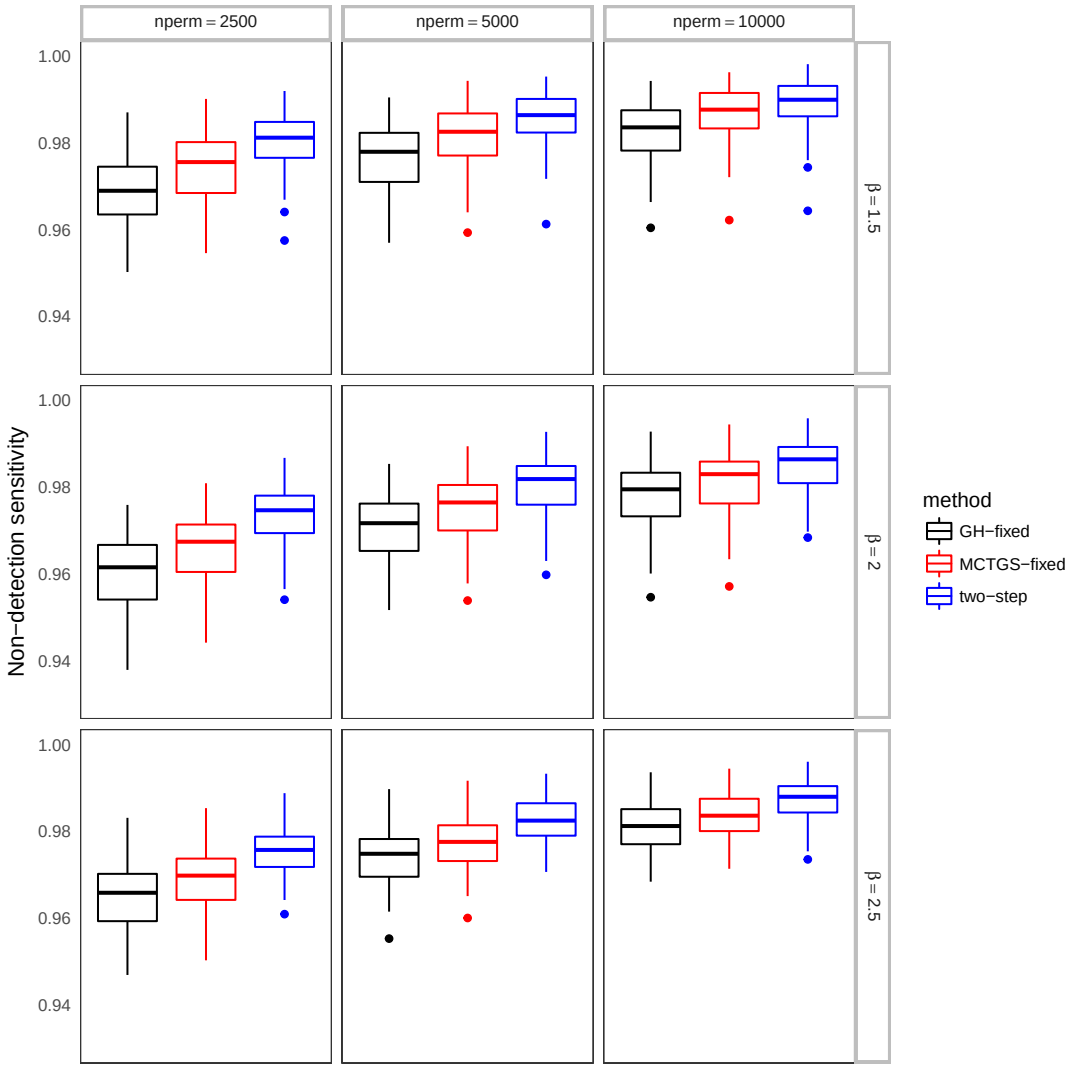

Figure B.5: The non-detection sensitivity in all 100 different realizations. The non-null ideal p-values were simulated from the Gaussian right-tailed model with $m = 1000, \beta = 1.5, 2$ and 2.5.

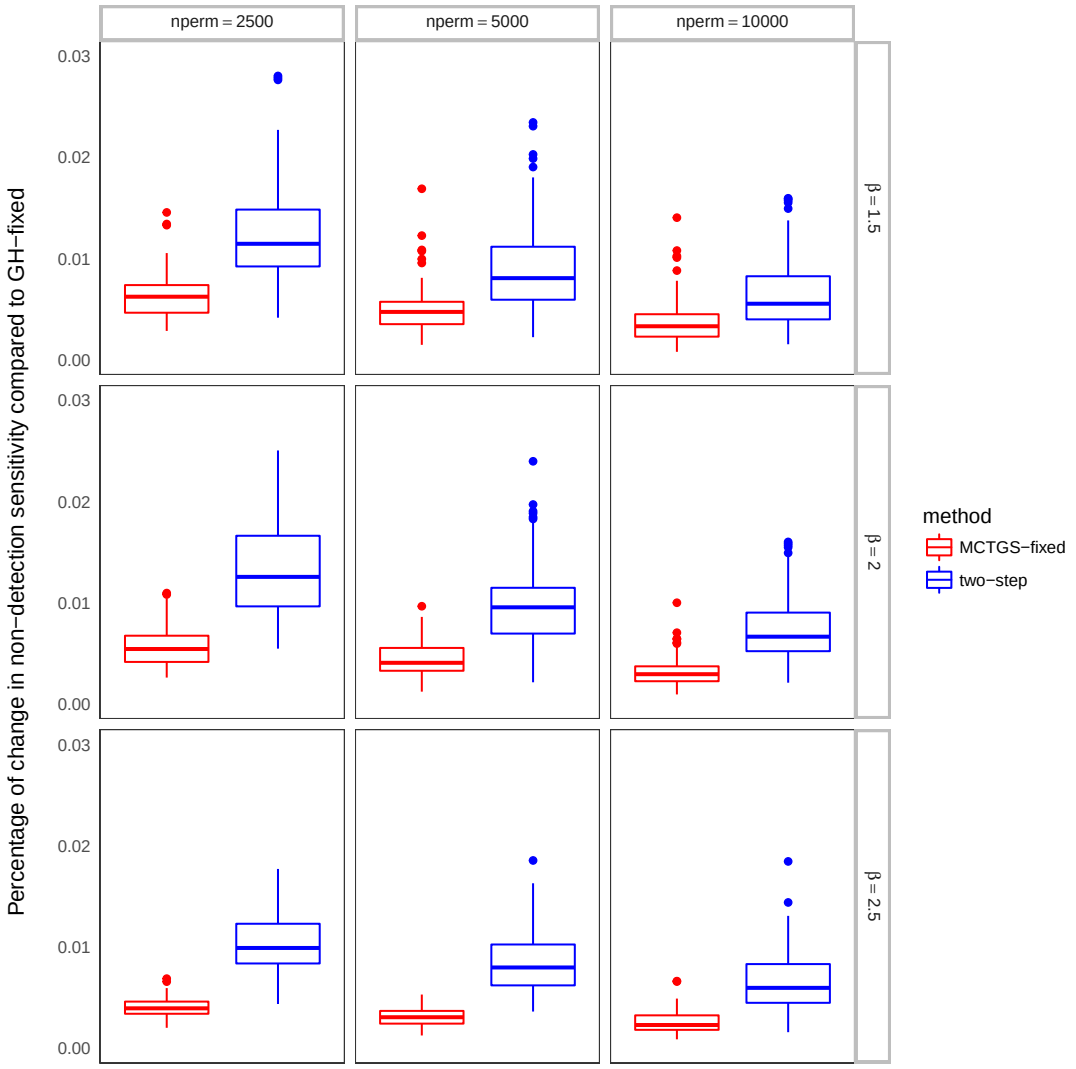

Figure B.6: The percentage of change in non-detection sensitivity compared to the GH-fixed method in all 100 different realizations. The non-null ideal p-values were simulated from the Gaussian right-tailed model with $m = 1000, \beta = 1.5, 2$ and 2.5.

Appendix C

Appendix for Chapter 4

C.1 Proof of Theorem 4.1

Proof of Theorem 4.1. We need to show that for any j, $\{|S_{i,j}(p)| < z\}$ is always an interval. let $z = \Phi^{-1}(1 - \alpha_i/4)$ Then $|S_{i,j}(p)| < z$ is equivalent to

$$f(p) = \left(\sum_{l=1}^{j} Z_{i,l} - (j \times N_{max}/G)p \right)^2 - z^2 N_{max} p(1-p) < 0.$$

It is easy to see that $f(p)$ is a quadratic polynomial of p and equation $f(p) = 0$ has at most two distinct roots. Consider $p_0 = (\sum_{l=1}^{j} Z_{i,l})/(j \times N_{max}/G)$. First because $0 \leq \sum_{l=1}^{j} Z_{i,l} \leq j \times N_{max}/G$, $0 \leq p_0 \leq 1$. The only occasion where $p_0 = 1$ is that $Z_{i,l} = N_{max}/G$ for every $l = 1, 2, \cdots, j$. The only occasion where $p_0 = 0$ is that $Z_{i,l} = 0$ for every $l = 1, 2, \cdots, j$. Excluding these two scenarios, $0 < p_0 < 1$ and

$$f(p_0) = 0 - z^2 N_{max} p_0 (1 - p_0) < 0.$$

At the same time, $f(0) = (\sum_{l=1}^{j} Z_{i,l})^2 > 0$ and $f(1) = (\sum_{l=1}^{j} Z_{i,l} - (j \times N_{max}/G))^2 > 0$. This further implies that there exist two distinct roots $0 < p_1 < p_2 < 1$ such that

$$f(p) \begin{cases} > 0, & p \in [0, p_1) \cup (p_2, 1] \\ = 0, & p = p_1 \text{ or } p = p_2 \\ < 0, & p \in (p_1, p_2) \end{cases}$$

Hence we know that the interval $\{|S_{i,j}(p)| < z\}$ can be written as (p_1, p_2). We also see that $\{S_{i,j}(p) < \Phi^{-1}(1-\alpha_i/4)\}$ can be written as $(p_1, 1)$ and that $\{S_{i,j}(p) > -\Phi^{-1}(1-\alpha_i/4)\}$ can be written as $(0, p_2)$. Next, when $p_0 = 1$,

$$f(p) = (j \times N_{max}/Gp)^2 - z^2 N_{max} p(1-p) = p\left((j \times N_{max}/G)^2 p - z^2 N_{max}(1-p)\right)$$
$$= ((j \times N_{max}/G)^2 + z^2 N_{max}) p \left(p - \frac{z^2 N_{max}}{(j \times N_{max}/G)^2 + z^2 N_{max}}\right).$$

The interval $\{|S_{i,j}(p)| < z\}$ can be now written as $(0, z^2 N_{max}/((j \times N_{max}/G)^2 + z^2 N_{max}) \subset (0,1)$. This happens to be the same interval for acceptance region inversion $\{S_{i,j}(p) > -\Phi^{-1}(1-\alpha_i/4)\}$. The other one-sided interval is $(0,1)$. Similarly we can show the result under $p_0 = 1$.

□

C.2 Proof of Theorem 4.2

Before proving the main results in Theorem 4.2, we first establish a connection between Algorithm 2 and a version using fixed number of replicates (Algorithm 5). Assume the final set of decisions based on algorithm 4 and 5 are $\mathbf{D}^{(2)} = (D_1^{(2)}, D_2^{(2)}, \cdots, D_m^{(2)})$ and $\mathbf{D}^{(3)} = (D_1^{(3)}, D_2^{(3)}, \cdots, D_m^{(3)})$ respectively. It is easy to see that $D_i^{(2)} = \textit{acceptance}$ implies $D_i^{(3)} = \textit{acceptance}$ and $D_i^{(2)} = \textit{rejection}$ implies $D_i^{(3)} = \textit{rejection}$, because the ons-sided (lower bound) sequential confidence at the end of group j

$$\bigcap_{l=1}^{j} \{S_{i,l}(p) < c_l(\alpha_i)\} \supset \bigcap_{l=1}^{G} \{S_{i,l}(p) < c_l(\alpha_i)\},$$

and similarly the other (upper bound) sequential interval

$$\bigcap_{l=1}^{j} \{S_{i,l}(p) > -c_l(\alpha_i)\} \supset \bigcap_{l=1}^{G} \{S_{i,l}(p) > -c_l(\alpha_i)\}.$$

In other words, the rejection and acceptance decisions with Algorithm 4 must be a subset of the rejection and acceptance decisions with Algorithm 5. Hence we only need to show the family-wise MCER with Algorithm 5 is controlled by α.

Proof of Theorem 4.2. Recall that the definition of family-wise MCER $= \Pr(V^{\mathrm{ME}} \geq 1)$,

Initialize $\widehat{D}_1, \widehat{D}_1, \cdots, \widehat{D}_m$ to be *undecided*.
Initialize $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ to be $(0, 1)$.
Initialize m_u to be m, and $m_{u'}$ to be $(m+1)$.

▷ Resampling begins

while $m_u \neq m_{u'}$ **do**
  $m_u = \#\{i \in \{1, 2, \cdots, m\} \mid \widehat{D}_i = \textit{undecided}\}$.
  $m_r = \#\{i \in \{1, 2, \cdots, m\} \mid \widehat{D}_i = \textit{rejection}\}$.
  $m_a = \#\{i \in \{1, 2, \cdots, m\} \mid \widehat{D}_i = \textit{acceptance}\}$.

  ▷ The sum $m_u + m_r + m_a$ is always m

  $m_{u'} = m_u$
  for *i in* $1 : m$ **do**
    if $\widehat{D}_i = \textit{undecided}$ **then**
      Find $\alpha^+ = \alpha/(m - m_a)$.
      Assign one-sided group sequential interval $\mathcal{A}_i^{\text{gs}+} = \bigcap_{j=1}^{G} \{S_{i,j}(p) < c_j(\alpha^+)\}$.
      Find $\alpha^- = \alpha/(m - m_r)$.
      Assign one-sided group sequential interval $\mathcal{A}_i^{\text{gs}-} = \bigcap_{j=1}^{G} \{S_{i,j}(p) > -c_j(\alpha^-)\}$.
      The two-sided group sequential interval $\mathcal{I}_i = \mathcal{A}_i^{\text{gs}+} \cap \mathcal{A}_i^{\text{gs}-}$.
    end
  end
  Apply standard BH on the upper limits of intervals $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ and decisions are $d_1^u, d_2^u, \cdots, d_m^u$.
  Apply standard BH on the lower limits of intervals $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m$ and decisions are $d_1^l, d_2^l, \cdots, d_m^l$.
  for *i in* $1 : m$ **do**
    if $d_i^u = d_i^l = \textit{acceptance}$ **then**
      Update $\widehat{D}_i =$*acceptance*.
    end
    if $d_i^u = d_i^l = \textit{rejection}$ **then**
      Update $\widehat{D}_i =$*rejection*.
    end
  end
end

▷ Resampling ends

return $(\widehat{D}_1, \widehat{D}_1, \cdots, \widehat{D}_m)$.

Algorithm 5: The MC tests with fixed number of replicates and group sequential intervals

where

$$
\begin{aligned}
V^{\text{MC}} &= \sum_{i=1}^{m} \left\{ \mathbb{I}(\widehat{D}_i = \textit{rejection}, D_i^* = \textit{acceptance}) + \mathbb{I}(\widehat{D}_i = \textit{acceptance}, D_i^* = \textit{rejection}) \right\}. \\
&= \sum_{i=1}^{m} \mathbb{I}(\widehat{D}_i = \textit{rejection}, D_i^* = \textit{acceptance}) + \sum_{i=1}^{m} \mathbb{I}(\widehat{D}_i = \textit{acceptance}, D_i^* = \textit{rejection}) \\
&= V^{\text{MC-I}} + V^{\text{MC-II}}.
\end{aligned}
$$

This implies that $\{V^{\text{MC}} \geq 1\} \subset \{V^{\text{MC-I}} \geq 1\} \cup \{V^{\text{MC-II}} \geq 1\}$ and $\Pr(V^{\text{MC}} \geq 1) \leq \Pr(V^{\text{MC-I}} \geq 1) + \Pr(V^{\text{MC-II}} \geq 1)$. The first term $\Pr(V^{\text{MC-I}} \geq 1)$ is exactly family-wise MCER-I. If we can show the Algorithm 5 controls $\Pr(V^{\text{MC-I}} \geq 1)$ and $\Pr(V^{\text{MC-II}} \geq 1)$ by $\alpha/2$ respectively, we prove the desired results.

Now we show the family-wise MCER-I, $\Pr(V^{\text{MC-I}} \geq 1)$ is bounded by $\alpha/2$. The other inequality can be proved with the same technique. For the i-th test, the null hypothesis is $H_i^{>} : p_i^*$ in BH's acceptance region and the alternative hypothesis is $H_i^{\leq} : p_i^*$ in BH's rejection region. The key of this proof is using the partition principle (Finner and Strassburger, 2002; Strassburger and Bretz, 2008) to create the one-sided (upper bound) compatible simultaneous confidence intervals for ideal p-values. We define such a partition of all possible ideal p-values $\mathbb{P}$ on the space $[0,1]^m$:

$$
\mathbb{P} = \left\{ P_{\mathscr{I}} \middle| P_{\mathscr{I}} = \{ i \in \mathscr{I}, p_i^* \text{ accepted by BH }, i \in \{1, 2, \cdots, m\}/\mathscr{I}, p_i^* \text{ rejected by BH } \} \right\}.
$$

Basically each element $P_{\mathscr{I}}$ corresponds to a region in $[0,1]^m$. In this region, p-values with indices in $\mathscr{I} \subset \{1, 2, \cdots, m\}$ are rejected by BH . Remaining p-values are accepted. Due to the uniqueness of decisions by BH, for two different $\mathscr{I}_1$ and $\mathscr{I}_2$ we have $P_{\mathscr{I}_1} \cap P_{\mathscr{I}_2} = \emptyset$. Therefore the partition $\mathbb{P}$ satisfies the disjoint assumption of the partition principle in Finner and Strassburger (2002). The compatible confidence intervals of p^* can be written as

$$
\mathscr{C} = \bigcup_{\mathscr{I}} \left\{ P_{\mathscr{I}} \bigcap_{i \in \mathscr{I}} \{ p_i^* \in \mathcal{A}_i^{\text{gs}-}(1 - \frac{\alpha}{2|\mathscr{I}|}) \} \bigcap_{i \notin \mathscr{I}} \{ p_i^* \in [0,1] \} \right\}. \tag{C1}
$$

Here we slightly modify the notation of $\mathcal{A}_i^{\text{gs}-}$, the sequential confidence intervals at the last

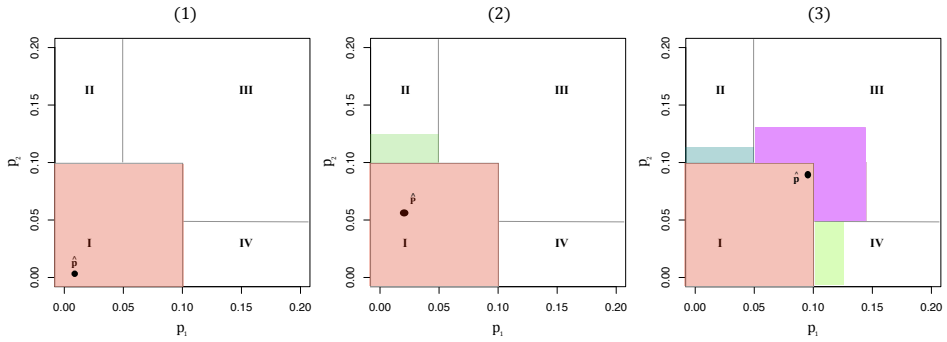

Figure C.1: An illustrative example (m = 2) of using partition principle for constructing simultaneous confidence intervals for (p_1^*, p_2^*) that are compatible with the decisions of standard BH procedure. The highlighted regions by different colors represent the local simultaneous confidence intervals within different regions from I to IV. In region I, both p_1^* and p_2^* should rejected by BH. In region II (IV), only p_1^* (p_2^*) should be rejected. In region III, none can be rejected. The union of all local intervals will be the proposed simultaneous upper bound intervals. The shape may vary a lot, depending on different combination of $(\widehat{p}_1, \widehat{p}_2)$.

group. Inside the parenthesis indicates the confidence level. Basically given the indices $\mathscr{I}$ for acceptance, set $\bigcap_{i \notin \mathscr{I}} \{p_i^* \in \mathcal{A}_i^{\text{gs}-}(1 - \frac{\alpha}{2|\mathscr{I}|})\}$ represents the simultaneous confidence intervals using the Bonferroni correction for those ideal p-values in the acceptance (null) region. Based on partition principle, we know that the coverage probability exceeds $(1 - \alpha/2)$, or in other words $\Pr(\mathbf{p}^* \in \mathscr{C}) \geq 1 - \alpha/2$. We claim a test to be rejected, only if by the BH rule every element within $\mathscr{C}$ rejects at this test. This approach must control family-wise MCER-I by α.

The final step is then to show that the union of rejections using Algorithm 5 is a subset of the rejections that are decided based on compatible confidence intervals $\mathscr{C}$. If this claim is not true, there must be an index $i^\dagger$ such that $\widehat{D}_{i^\dagger} = \textit{rejection}$ using Algorithm 3, but there exists an $\mathscr{I}$ such that $i^\dagger \in \mathscr{I}$ and

$$P_{\mathscr{I}} \bigcap_{i \in \mathscr{I}} \{p_i^* \in \mathcal{A}_i^{\text{gs}-}(1 - \frac{\alpha}{2|\mathscr{I}|})\} \neq \emptyset. \tag{C2}$$

The indices of all p_i^* in the acceptance region are denoted by $\mathscr{I}^*$. It can be shown that $|\mathscr{I}| \leq |\mathscr{I}^*|$, otherwise the LHS in (C2) must be an empty set. However, when $|\mathscr{I}| \leq |\mathscr{I}^*|$, we know that $\mathcal{A}_{i^\dagger}^{\text{gs}-}(1 - \frac{\alpha}{2|\mathscr{I}|}) \subset \mathcal{A}_{i^\dagger}^{\text{gs}-}(1 - \frac{\alpha}{2|\mathscr{I}^*|})$. Because $\widehat{D}_{i^\dagger} = \textit{rejection}$, then both on-sided intervals must be in the rejection region of BH. This contradicts with the fact that $i^\dagger$ is an index of the acceptance region. □

Bibliography

Alishahi, K., Ehyaei, A. R., and Shojaie, A. (2016), "A generalized benjamini-hochberg procedure for multivariate hypothesis testing," *arXiv preprint arXiv:1606.02386*, .

Anderson, M. J. (2001), "A new method for non-parametric multivariate analysis of variance," *Austral Ecology*, 26(1), 32–46.

Armitage, P. (1958), "Sequential methods in clinical trials," *American Journal of Public Health and the Nations Health*, 48(10), 1395–1402.

Baker, M. (2016), "1,500 scientists lift the lid on reproducibility," *Nature News*, 533(7604), 452.

Barber, R. F., and Candès, E. J. (2015), "Controlling the false discovery rate via knockoffs," *The Annals of Statistics*, 43(5).

Barber, R. F., and Ramdas, A. (2017), "The p-filter: multilayer false discovery rate control for grouped hypotheses," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4), 1247–1268.

Bauer, P., and Kohne, K. (1994), "Evaluation of experiments with adaptive interim analyses," *Biometrics*, pp. 1029–1041.

Bauer, P., Röhmel, J., Maurer, W., and Hothorn, L. (1998), "Testing strategies in multi-dose experiments including active control," *Statistics in Medicine*, 17(18), 2133–2146.

Benjamini, Y., and Bogomolov, M. (2014), "Selective inference on multiple families of hypotheses," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 297–318.

Benjamini, Y., and Heller, R. (2007), "False discovery rates for spatial signals," *Journal of the American Statistical Association*, 102(480), 1272–1281.

Benjamini, Y., and Heller, R. (2008), "Screening for partial conjunction hypotheses," *Biometrics*, 64(4), 1215–1222.

Benjamini, Y., Heller, R., and Yekutieli, D. (2009), "Selective inference in complex research," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906), 4255–4271.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Benjamini, Y., and Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, 29(4), 1165–1188.

Besag, J., and Clifford, P. (1991), "Sequential Monte Carlo p-values," *Biometrika*, 78(2), 301–304.

Bonferroni, C. (1936), "Teoria statistica delle classi e calcolo delle probabilita," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8, 3–62.

Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009), "A graphical approach to sequentially rejective multiple test procedures," *Statistics in medicine*, 28(4), 586–604.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001), "Interval estimation for a binomial proportion," *Statistical Science*, 16(2), 101–133.

Brown, L. D., Cai, T. T., and DasGupta, A. (2002), "Confidence intervals for a binomial proportion and asymptotic expansions," *The Annals of Statistics*, 30(1), 160–201.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3), 551–577.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996), "Asymptotics for the SIMEX estimator in nonlinear measurement error models," *Journal of the American Statistical Association*, 91(433), 242–250.

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007), "A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria," *Journal of microbiological methods*, 69(2), 330–339.

Chen, E. Z., and Li, H. (2016), "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data," *Bioinformatics*, 32(17), 2611–2617.

Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014), "GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation," *PLoS genetics*, 10(11), e1004787.

Coe, P. R., and Tamhane, A. C. (1993), "Exact repeated confidence intervals for Bernoulli parameters in a group sequential clinical trial," *Controlled clinical trials*, 14(1), 19–29.

Darling, D., and Robbins, H. (1968), "Some further remarks on inequalities for sample sums," *Proceedings of the National Academy of Sciences of the United States of America*, 60(4), 1175.

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap methods and their application*, Vol. 1 Cambridge university press.

Davison, A. C., Hinkley, D. V., and Young, G. A. (2003), "Recent developments in bootstrap methodology," *Statistical Science*, pp. 141–157.

Davison, A., Hinkley, D. V., and Schechtman, E. (1986), "Efficient bootstrap simulation," *Biometrika*, 73(3), 555–566.

De Finetti, B. (1992), "Foresight: Its logical laws, its subjective sources," in *Breakthroughs in statistics* Springer, pp. 134–174.

Demets, D. L., and Lan, K. G. (1994), "Interim analysis: the alpha spending function approach," *Statistics in medicine*, 13(13-14), 1341–1352.

Dobriban, E. (2018), "Flexible multiple testing with the FACT algorithm," *arXiv preprint arXiv:1806.10163*, .

Doob, J. L. (1953), *Stochastic processes*, Vol. 101 Wiley, New York.

Du, L., Zhang, C. et al. (2014), "Single-index modulated multiple testing," *The Annals of Statistics*, 42(4), 1262–1311.

Durrett, R. (2019), *Probability: theory and examples*, Vol. 49 Cambridge university press.

Efron, B. (1987), "Better bootstrap confidence intervals," *Journal of the American statistical Association*, 82(397), 171–185.

Efron, B. (2007), "Correlation and large-scale simultaneous significance testing," *Journal of the American Statistical Association*, 102(477), 93–103.

Efron, B. (2012), *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1 Cambridge University Press.

Efron, B. (2016), "Empirical Bayes deconvolution estimates," *Biometrika*, 103(1), 1–20.

Efron, B., and Tibshirani, R. (1998), "The problem of regions," *The Annals of Statistics*, 26(5), 1687–1718.

Fan, J., Han, X., and Gu, W. (2012), "Estimating false discovery proportion under arbitrary covariance dependence," *Journal of the American Statistical Association*, 107(499), 1019–1035.

Fang, R., Wagner, B., Harris, J., and Fillon, S. (2016), "Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis," *Epidemiology & Infection*, 144(11), 2447–2455.

Farcomeni, A. (2007), "Some results on the control of the false discovery rate under dependence," *Scandinavian Journal of Statistics*, 34(2), 275–297.

Fay, M. P., and Follmann, D. A. (2002), "Designing Monte Carlo implementations of permutation or bootstrap hypothesis tests," *The American Statistician*, 56(1), 63–70.

Fidler, F., and Wilcox, J. (2018), "Reproducibility of scientific results," in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta, Winter 2018 edn Metaphysics Research Lab, Stanford University.

Finner, H., and Strassburger, K. (2002), "The partitioning principle: a powerful tool in multiple decision theory," *Annals of Statistics*, pp. 1194–1213.

Fisher, R. A. (1992), "Statistical methods for research workers," in *Breakthroughs in Statistics* Springer, pp. 66–70.

Friguet, C., Kloareg, M., and Causeur, D. (2009), "A factor model approach to multiple testing under dependence," *Journal of the American Statistical Association*, 104(488), 1406–1415.

Friston, K. J., Penny, W. D., and Glaser, D. E. (2005), "Conjunction revisited," *Neuroimage*, 25(3), 661–667.

Gandy, A. (2009), "Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk," *Journal of the American Statistical Association*, 104(488), 1504–1511.

Gandy, A., and Hahn, G. (2014), "MMCTest—a safe algorithm for implementing multiple Monte Carlo tests," *Scandinavian Journal of Statistics*, 41(4), 1083–1101.

Gandy, A., and Hahn, G. (2016), "A framework for Monte Carlo based multiple testing," *Scandinavian Journal of Statistics*, 43(4), 1046–1063.

Gavrilov, Y., Benjamini, Y., and Sarkar, S. K. (2009), "An adaptive step-down procedure with proven FDR control under independence," *The Annals of Statistics*, 37(2), 619–629.

Genovese, C., and Wasserman, L. (2004), "A stochastic process approach to false discovery control," *The Annals of Statistics*, 32(3), 1035–1061.

Genz, A., and Bretz, F. (2009), *Computation of multivariate normal and t probabilities*, Vol. 195 Springer Science & Business Media.

Gleason, J. R. (1988), "Algorithms for balanced bootstrap simulations," *The American Statistician*, 42(4), 263–266.

Gordon Lan, K., and DeMets, D. L. (1983), "Discrete sequential boundaries for clinical trials," *Biometrika*, 70(3), 659–663.

Guo, W., and Peddada, S. (2008), "Adaptive choice of the number of bootstrap samples in large scale multiple testing," *Statistical applications in genetics and molecular biology*, 7(1).

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'Amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., McClure, E. E., Dunklebarger, M. F., Knight, R., and Jansson, J. K. (2017), "Dynamics of the human gut microbiome in inflammatory bowel disease," *Nature Microbiology*, 2(5), 17004.

Hall, P. (1986), "On the number of bootstrap simulations required to construct a confidence interval," *The Annals of Statistics*, 14(4), 1453–1462.

Hansen, P. R. (2005), "A test for superior predictive ability," *Journal of Business & Economic Statistics*, 23(4), 365–380.

Heller, R., Chatterjee, N., Krieger, A., and Shi, J. (2018), "Post-selection inference following aggregate level hypothesis testing in large-scale genomic data," *Journal of the American Statistical Association*, 113(524), 1770–1783.

Hinkley, D. V. (1988), "Bootstrap methods," *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3), 321–337.

Holm, S. (1979), "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 6(2), 65–70.

Hommel, G., Bretz, F., and Maurer, W. (2007), "Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies," *Statistics in Medicine*, 26(22), 4063–4073.

Hsiao, E. Y., McBride, S. W., Hsien, S., Sharon, G., Hyde, E. R., McCue, T., Codelli, J. A., Chow, J., Reisman, S. E., Petrosino, J. F. et al. (2013), "Microbiota modulate

behavioral and physiological abnormalities associated with neurodevelopmental disorders," *Cell*, 155(7), 1451–1463.

Hsu, P.-H., Hsu, Y.-C., and Kuan, C.-M. (2010), "Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias," *Journal of Empirical Finance*, 17(3), 471–484.

Hu, Y., and Satten, G. A. (2017), "Testing hypotheses about the microbiome using an ordination-based linear decomposition model," *BioRxiv*, p. 229831.

Javanmard, A., and Montanari, A. (2018), "Online rules for control of false discovery rate and false discovery exceedance," *The Annals of statistics*, 46(2), 526–554.

Jiang, H., and Salzman, J. (2012), "Statistical properties of an early stopping rule for resampling-based multiple testing," *Biometrika*, 99(4), 973–980.

Khintchine, A. (1924), "über einen satz der wahrscheinlichkeitsrechnung," *Fundamenta Mathematicae*, 6(1), 9–20.

Kim, H.-J. (2010), "Bounding the resampling risk for sequential Monte Carlo implementation of hypothesis tests," *Journal of Statistical Planning and Inference*, 140(7), 1834–1843.

Kimmel, G., and Shamir, R. (2006), "A fast method for computing high-significance disease association in large population-based studies," *The American Journal of Human Genetics*, 79(3), 481–492.

Koh, H., Blaser, M. J., and Li, H. (2017), "A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping," *Microbiome*, 5(1), 45.

Lai, T. L. (1976), "On confidence sequences," *The Annals of Statistics*, 4, 265–280.

Leek, J. T., and Storey, J. D. (2008), "A general framework for multiple testing dependence," *Proceedings of the National Academy of Sciences*, 105(48), 18718–18723.

Lehmacher, W., and Wassmer, G. (1999), "Adaptive sample size calculations in group sequential trials," *Biometrics*, 55(4), 1286–1290.

Lehmann, E. L., and Romano, J. P. (2006), *Testing statistical hypotheses* Springer Science & Business Media.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012), "Normalization, testing, and false discovery rate estimation for RNA-sequencing data," *Biostatistics*, 13(3), 523–538.

Liu, R. Y., and Singh, K. (1997), "Notions of limiting P values based on data depth and bootstrap," *Journal of the American Statistical Association*, 92(437), 266–277.

Liu, Y., and Xie, J. (2019), "Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures," *Journal of the American Statistical Association*, pp. 1–18.

Loughin, T. M. (2004), "A systematic comparison of methods for combining p-values from independent tests," *Computational statistics & data analysis*, 47(3), 467–485.

Love, M. I., Huber, W., and Anders, S. (2014), "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, 15, 550.

Marcus, R., Eric, P., and Gabriel, K. R. (1976), "On closed testing procedures with special reference to ordered analysis of variance," *Biometrika*, 63(3), 655–660.

Mehta, C. R., Bauer, P., Posch, M., and Brannath, W. (2007), "Repeated confidence intervals for adaptive group sequential trials," *Statistics in Medicine*, 26(30), 5422–5433.

Meinshausen, N. (2008), "Hierarchical testing of variable importance," *Biometrika*, 95(2), 265–278.

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B. et al. (2012), "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment," *Genome Biology*, 13(9), R79.

Müller, H.-H., and Schäfer, H. (2001), "Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches," *Biometrics*, 57(3), 886–891.

Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J.-B. (2005), "Valid conjunction inference with the minimum statistic," *NeuroImage*, 25, 653–660.

O'Brien, P. C., and Fleming, T. R. (1979), "A multiple testing procedure for clinical trials," *Biometrics*, pp. 549–556.

Owen, A. B. (2009), "Karl Pearson's meta-analysis revisited," *The Annals of Statistics*, 37(6B), 3867–3892.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013), "Differential abundance analysis for microbial marker-gene surveys," *Nature Methods*, 10(12), 1200–1202.

Pearson, E. S. (1938), "The probability integral transformation for testing goodness of fit and combining independent tests of significance," *Biometrika*, 30(1/2), 134–148.

Peng, R. D. (2011), "Reproducible research in computational science," *Science*, 334(6060), 1226–1227.

Peterson, C. B., Bogomolov, M., Benjamini, Y., and Sabatti, C. (2016), "TreeQTL: hierarchical error control for eQTL findings," *Bioinformatics*, 32(16), 2556–2558.

Phipson, B., and Smyth, G. K. (2010), "Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn," *Statistical applications in genetics and molecular biology*, 9(1).

Pillai, N. S., and Meng, X. L. (2016), "An unexpected encounter with Cauchy and Lévy," *The Annals of Statistics*, 44(5), 2089–2097.

Pocock, S. J. (1977), "Group sequential methods in the design and analysis of clinical trials," *Biometrika*, 64(2), 191–199.

Pounds, S., and Morris, S. W. (2003), "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values," *Bioinformatics*, 19(10), 1236–1242.

Price, C. J., and Friston, K. J. (1997), "Cognitive conjunction: a new approach to brain activation experiments," *Neuroimage*, 5(4), 261–270.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., and Wang, J. (2010), "A human gut microbial gene catalog established by metagenomic sequencing," *Nature*, 464(7285), 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D. et al. (2012), "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, 490(7418), 55.

Qiu, X., Klebanov, L., and Yakovlev, A. (2005), "Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes," *Statistical applications in genetics and molecular biology*, 4(1).

Ramdas, A., Chen, J., Wainwright, M. J., and Jordan, M. I. (2017), "DAGGER: A sequential algorithm for FDR control on DAGs," *arXiv preprint arXiv:1709.10250*, .

Robbins, H. (1970), "Statistical methods related to the law of the iterated logarithm," *The Annals of Mathematical Statistics*, 41(5), 1397–1409.

Robbins, H., and Siegmund, D. (1972), A class of stopping rules for testing parametric hypotheses,, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Health*, The Regents of the University of California, The Regents of the University of California.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010), "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, 26, 139–140.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008), "Control of the false discovery rate under dependence using the bootstrap and subsampling," *Test*, 17(3), 417.

Romano, J. P., and Wolf, M. (2005), "Exact and approximate stepdown methods for multiple hypothesis testing," *Journal of the American Statistical Association*, 100(469), 94–108.

Rosenbaum, P. R. (2008), "Testing hypotheses in order," *Biometrika*, 95(1), 248–252.

Sandve, G. K., Ferkingstad, E., and Nygård, S. (2011), "Sequential Monte Carlo multiple testing," *Bioinformatics*, 27(23), 3235–3241.

Sankaran, K., and Holmes, S. (2014), "structSSI: simultaneous and selective inference for grouped or hierarchically structured data," *Journal of statistical software*, 59(13), 1–21.

Schwartzman, A., and Lin, X. (2011), "The effect of correlation in false discovery rate estimation," *Biometrika*, 98(1), 199–214.

Shaffer, J. P. (1980), "Control of directional errors with stagewise multiple test procedures," *The Annals of Statistics*, 8(6), 1342–1347.

Shimodaira, H. (2008), "Testing regions with nonsmooth boundaries via multiscale bootstrap," *Journal of Statistical Planning and Inference*, 138(5), 1227–1241.

Simes, R. J. (1986), "An improved Bonferroni procedure for multiple tests of significance," *Biometrika*, 73(3), 751–754.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. et al. (2002), "Gene expression correlates of clinical prostate cancer behavior," *Cancer cell*, 1(2), 203–209.

Stefanski, L. A., and Cook, J. R. (1995), "Simulation-extrapolation: the measurement error jackknife," *Journal of the American Statistical Association*, 90(432), 1247–1256.

Storey, J. D. (2002), "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 479–498.

Storey, J. D. (2003), "The positive false discovery rate: a Bayesian interpretation and the q-value," *The Annals of Statistics*, 31(6), 2013–2035.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified

approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1), 187–205.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. (1949), "The American soldier: Adjustment during army life.(Studies in social psychology in World War II), Vol. 1," , .

Strassburger, K., and Bretz, F. (2008), "Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests," *Statistics in medicine*, 27(24), 4914–4927.

Sun, W., and Cai, T. T. (2009), "Large-scale multiple testing under dependence," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2), 393–424.

Tamhane, A. C., and Liu, L. (2008), "On weighted Hochberg procedures," *Biometrika*, 95(2), 279–294.

Tang, W. W., and Hazen, S. L. (2017), "The gut microbiome and its role in cardiovascular diseases," *Circulation*, 135(11), 1008–1010.

Tang, Y., Ma, L., and Nicolae, D. L. (2018), "A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data," *The Annals of Applied Statistics*, 12(1), 1–26.

Tang, Z.-Z., Chen, G., Alekseyenko, A. V., and Li, H. (2017), "A general framework for association analysis of microbial communities on a taxonomic tree," *Bioinformatics*, 33(9), 1278–1285.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. I. (2007), "The human microbiome project: exploring the microbial part of ourselves in a changing world," *Nature*, 449(7164), 804–810.

Wald, A. (1945), "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, 16(2), 117–186.

Wilson, E. B. (1927), "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, 22(158), 209–212.

Wu, W. B. (2008), "On false discovery control under dependence," *The Annals of statistics*, 36(1), 364–380.

Yekutieli, D. (2008), "Hierarchical false discovery rate–controlling methodology," *Journal of the American Statistical Association*, 103(481), 309–316.

Yu, K., Liang, F., Ciampa, J., and Chatterjee, N. (2011), "Efficient p-value evaluation for resampling-based tests," *Biostatistics*, 12(3), 582–593.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015), "Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test," *The American Journal of Human Genetics*, 96(5), 797–807.

Zhao, S. D. (2015), "False discovery rate control for identifying simultaneous signals," *arXiv preprint arXiv:1512.04499*, .