Multiple Testing for Data with Ancillary Information

by

Xuechan Li

Department of Biostatistics and Bioinformatics
Duke University

Date: _____

Approved:

_____
Jichun Xie, Supervisor

_____
Cliburn Chan

_____
Josh Granek

_____
Kouros Owzar

_____
Li Ma

_____
Yongtao Guan

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Biostatistics and Bioinformatics
in the Graduate School of Duke University
2022

Abstract

Multiple Testing for Data with Ancillary Information

by

Xuechan Li

Department of Biostatistics and Bioinformatics
Duke University

Date: _____

Approved:

_____
Jichun Xie, Supervisor

_____
Cliburn Chan

_____
Josh Granek

_____
Kouros Owzar

_____
Li Ma

_____
Yongtao Guan

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics and Bioinformatics
in the Graduate School of Duke University
2022

Abstract

In my dissertation, I develop three powerful hierarchical multiple testing methods by accounting for ancillary information of data. In my first project, we develop a multiple testing framework named Distance Assisted Recursive Testing (DART). DART assumes there exists some informative distance information in the data. Through rigorous proof and extensive simulations, we justified the false discovery rate (FDR) control and sensitivity improvement of DART. As an illustration, we apply our method to a clinical trial in leukemia patients receiving hematopoietic cell transplantation to identify the gut microbiota whose abundance will be impacted by the after-transplant care. The second project is motivated by the flow cytometry analysis in immunology study. The analysis can be translated into a statistical problem which is trying to pinpoint the regions where two density functions differ. By partitioning the sample space into small bins and conducting testing on each bin, we model the analysis into a multiple testing problem. We provide theoretical justification that the procedure achieves the statistical goal of pinpointing the regions with differential density with high sensitivity and precision. My third project is motivated by the rare variant association study. We develop a multiple testing framework named DATED (Dynamic Aggregation and Tree-Embedded testing) to pinpoint the disease-associated rare-variant regions hierarchically and dynamically. To accommodate the application objective, DATED adopts a rare variant region-level FDR weighted by the proportions of the neutral rare-variant. Extensive numerical simulations demonstrate the

superior performance of DATED under various scenarios compared to the existing methods. We illustrate DATED by applying it to an amyotrophic lateral sclerosis (ALS) study for identifying pathogenic rare variant regions.

To my family

Contents

List of Tables

List of Figures

Acknowledgements

I am very fortunate to have spent the past years in the Department of Biostatistics and Bioinformatics at Duke University. It is a valuable memory and I would like to express my sincere appreciation to all faculties, students, and staff members.

First of all, I want to thank my advisor, Dr. Jichun Xie, for being a great mentor during my Ph.D. studies. Jichun has always been available whenever I needed support or mentorship. Her instructions and warm encouragement have helped me become a researcher and statistician. I would have hardly achieved what I have without her support.

Many thanks to the other five committee members Drs. Kouros Owzar, Cliburn Chan, Li Ma, Grant Guan, and Josh Granek for their time and guidance. I also want to thank my collaborator, Dr. John Pura. John collaborated with me on chapters 2 and 3.

I have received tremendous support from Drs. Kouros Owzar and Terry Hyslop. Kouros was my master thesis advisor and led me into Statistics. Terry was my supervisor when I worked at the Duke Cancer Institute. Their advice, encouragement, and support has helped to shape my career path, and me overcome failures and difficulties.

I appreciate the research assistantships with Dr. Mustafa Khasraw. These collaborative research projects have exposed me to important scientific questions and

interesting real data problems.

I wish to thank my lab mates Jiyuan Fang, Qi Gao, and Xiaodi Qin. I am very appreciative for their friendship and camaraderie. I would also like to thank the staff in our department, Ellen Baker, Shannon Clarke, Michelle Evans, Ladonna Huseman, Kendall Mincey, Kim Hall and Carolyn Walters for all their support over the years.

Finally, I would like to thank my parent and other family members for their unconditional love and support. I am very fortunate to have them in my life.

1

Introduction

In many applications, a large number of features are collected with the goal of identifying a few important ones among them. Sometimes, the important features tend to cluster based on some ancillary information such as distance information or localized structure within the data. Proper use of the ancillary information will boost the power of identifying important features. Hence, motivated by three different biomedical research problems, we develop three powerful hierarchical multiple testing frameworks by considering the distance information or localized structure within the data.

In my first project, we develop a general multiple testing framework named Distance Assisted Recursive Testing (DART) (Li et al., 2021). DART has two stages. In stage 1, we transform the distance matrix into an aggregation tree, where each node represents a set of features. In stage 2, based on the aggregation tree, we set up dynamic node hypotheses and perform multiple testing on the tree. All rejections are mapped back to the features. Under mild assumptions, the false discovery proportion of DART converges to the desired level in high probability converging to one. We

illustrate by theory and simulations that DART has superior performance under various models compared to the existing methods. We apply DART to a clinical trial of hematopoietic stem cell transplantation patients to identify gut microbiota whose abundance is impacted by the after-transplant care.

My second project is a collaborative work with Dr. John Pura. The second project develops a novel multiple testing method, called TEAM (Testing on the Aggregation tree Method), to pinpoint those regions that harbor differential probability density functions (PDFs) with application in flow cytometry analysis (Pura et al., 2019). With rigorous mathematical proof, I provide the theoretical justification that TEAM can provide powerful discoveries while controlling the false discovery rate (FDR) under the desired level.

The third project develops a novel analysis framework named DATED (Dynamic Aggemation and Tree-Embedded testing) to pinpoint the disease-associated rare variant sets with a controlled false discovery rate. Given the lack of sensitivity of classical single-variant association analysis on the rare genetic variants, researchers usually study the collective effect of sets of rare variants in dynamic or prefixed genetic regions. However, the existing methods either falsely detect too many neutral variants or fail to accommodate the unevenly distributed disease-associated variants. Unlike the existing methods, DATED dynamically and hierarchically aggregates smaller rare variant sets to larger ones and performs multiple testing on them. As a result, DATED has high sensitivity and precision. By conducting extensive numerical simulations, I demonstrate the superior performance of DATED under various scenarios compared to the existing methods. I also illustrate DATED by applying it to an amyotrophic lateral sclerosis (ALS) study for identifying pathogenic rare variants. This project is an extension of Dr. John Pura's previous work Pura (2020). I modify the original framework by adding a correction factor in the procedure for a more

robust FDR control. I also design a sampling algorithm for the estimation of the correction factor.

2

DART: Distance Assisted Recursive Testing

2.1 Introduction

A typical multiple testing problem aims to identify a small number of important features among many with a controlled false discovery rate. Sometimes, these features lie in a metric space with known pairwise distances. For example, in neuro-imaging studies, the distance between two neurons can be calculated based on their 3D location and the brain anatomy structure; in microbiome studies, the distance between any two amplicon sequence variants (ASVs) can be calculated based on their evolutionary distance; and in spatial analysis, the Euclidean distances between two sites can be calculated via their geometric locations. In these examples, neurons, ASVs, and geometric locations are features of interest. Very often, important features tend to cluster with each other. If two features are close in distance, they are likely to be co-important or co-unimportant. For example, in microbiome studies, two evolutionarily close ASVs often perform similar biological functions. If one is important, the other is probably important too. Thus when testing the ASV abundance association with the treatment, if we can properly incorporate their evolutionary distance,

the testing power will be boosted. In this chapter, we develop a new multiple testing method incorporating the distance information to boost the testing power while controlling the asymptotic feature-level false discovery rate (FDR).

The existing literature provides alternative solutions to incorporate distance information into testing. One of them is to model the features by hidden Markov chains (Sun and Cai, 2009) or hidden Markov random fields (Liu et al., 2012; Shu et al., 2015; Lee and Lee, 2016). The co-importance patterns are introduced by the transition probabilities between the importance and unimportance status among those features. The challenge lies in how to accurately inferring the transition probabilities. Even assuming all the feature statistics follow multivariate Gaussian distribution, it is still hard to derive consistent transition probability estimators without additional information. Another solution is to use the weighted or smoothed P-values in the neighborhood. Zhang et al. (2011) developed a method called FDR_L. FDR_L prespecified a smoothing window. For each hypothesis, it smooths the P-value across its local neighbors within the window. Recently, Cai et al. (2020) developed a locally-adaptive weighting and screening method named LAWS. LAWS weights the P-values using the estimated local sparsity level, which is calculated based on a pre-specified kernel function. However, the performance of LAWS heavily depends on the accuracy in local sparsity level estimation, which is difficult without additional information. In addition, LAWS focuses on a setting that the features are located in a regular lattice and require a non-vanishing proportion of features to be important. These conditions might not hold for many large-scale feature selection problems.

In this chapter, we propose a new solution called Distance Assisted Recursive Testing (DART). It embeds multiple testing into an aggregation tree built upon the feature distances. DART has two stages.

- Stage I is to construct an aggregation tree based on the distance matrix. First,

on layer 1, each node contains only one feature; it is also called a leaf. On layer ℓ ($\ell \geq 2$), we gradually aggregate the close child nodes from the previous layers to form new nodes on the current layer. The detailed algorithm is described in Section 2.2.2.

- Stage II is to perform multiple testing (of testing feature importance) on the aggregation tree from Stage I. On layer 1, we apply the multiple testing procedure to asymptotically control the feature-level FDR. Traditional multiple testing method will stop after one-layer of testing but DART will not. On layer ℓ ($\ell \geq 2$), the already-rejected child nodes from the previous layers will be excluded from the nodes on the current layer to form dynamic working nodes. Next, we apply the new multiple testing procedure on the working nodes to control the node-level FDR up to layer ℓ. If a node on layer ℓ is rejected then all its containing features will be rejected. This rejection rule is very aggressive but the feature-level FDR will still be asymptotically controlled under mild conditions (See Section 2.3). The detailed algorithm is described in Section 2.2.2.

The underlying logic of DART lies in the assumption that closer features are more likely to have co-importance or co-unimportance patterns. Some important features could have weak signal-to-noise ratios. If one such feature stands alone, its chance to be discovered is hampered by the weak signal-to-noise ratios; if several such features are aggregated, their collective signal-to-noise ratios will be amplified, and thus their chances to be discovered are boosted.

The rest of the chapter is organized as follows. Section 2.2 describes the DART algorithm. Section 2.3 justifies the asymptotic validity of DART under mild conditions. Section 2.4 shows that under various models, DART has superior performance than competing methods. Section 2.5 applies DART to study the impact hematopoietic

stem cell transplantation (HCT) post-transplant care on patient gut microbiota compositions. Section 2.6 provides a brief discussion on the possible extension of DART. The proofs of propositions and theorems are provided in Section 2.7. More details on the DART algorithms and the proofs of the lemmas are provided in the Appendix A.

2.2 Method

2.2.1 Model

Denote by $\Omega = \{1, \ldots, m\}$ the set with m features. Assume the distance matrix of these features is $\boldsymbol{D} = (d_{ij})_{m\times m}$, where $d_{ij} = d_{ji}$ is the distance between feature i and feature j. We define $d_{ii} = 0$. The distance matrix can be scaled so that $\max_{i\neq j} d_{ij} = 1$.

Among these features, let Ω_1 be the important (alternative) feature set, Ω_0 be the unimportant (null) feature set, and $\Omega_1 \cap \Omega_0 = \emptyset$, $\Omega_1 \cup \Omega_0 = \Omega$. For feature i, the hypothesis is

$$\mathrm{H}_{0i} : i \in \Omega_0 \quad \text{versus} \quad \mathrm{H}_{1i} : i \in \Omega_1. \tag{2.1}$$

To test $\mathrm{H}_{0,i}$, a feature P-value (statistic) T_i is derived.

Definition 2.1 (Oracle P-value)**.** We call a statistic $\tilde{T}_i$ an oracle P-value if

$$\mathsf{P}(\tilde{T}_i \leq p) \leq p \text{ when } i \in \Omega_0 \quad \text{and} \quad \mathsf{P}(\tilde{T}_i \leq p) > p \text{ when } i \in \Omega_1.$$

Under many circumstances, the P-values are derived from asymptotic tests (such as Wald test, the score test, and the likelihood ratio test), and thus are not oracle P-values; however, they asymptotically converge to the oracle P-values.

Definition 2.2 (Asymptotic oracle P-value)**.** We call a statistic T_i an asymptotic oracle P-value if

$$\sup_{i\in\Omega_0}\sup_{p\in\mathcal{P}_{i0}}\left|\frac{\mathsf{P}(T_i < p)}{\mathsf{P}(\tilde{T}_i < p)} - 1\right| \leq \delta_{0m} \quad \text{with} \lim_{m\to\infty} \delta_{0m} = o(1), \tag{2.2}$$

where $\mathcal{P}_{i0} = \left\{p \in [0,1] : P(\tilde{T}_i < p) \geq \left\{m(\log m \log\log m)^{1/2}\right\}^{-1}\right\}$.

In this study, we assume all the feature P-values are asymptotic oracle P-values. This assumption is easily satisfied by many commonly used models and tests. Here we provide a linear model example with features as outcomes, which is applied to study the impact of HCT post-transplant care on patient gut microbiota complications. Please see Section 2.5 for details.

Example 2.1. Consider the linear regression model:

$$\boldsymbol{Y}_{n\times m} = \boldsymbol{W}_{n\times p_0}\boldsymbol{\theta}_{p_0\times m} + \boldsymbol{\epsilon}_{n\times m}, \tag{2.3}$$

where $\boldsymbol{Y}_{n\times m} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)$ is a feature outcome matrix with n observations of m features allowing $m > n$, $\boldsymbol{W}_{n\times p_0}$ is the design matrix with n observations of p_0 covariants with $p_0 < n$, $\boldsymbol{\theta}_{p_0\times m} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ is the coefficient matrix with $\boldsymbol{\theta}_i \in \mathbb{R}^{p_0}$ the coefficient of $\boldsymbol{W}$ on $\boldsymbol{Y}_i$, and $\boldsymbol{\epsilon}_{n\times m} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_m)$ is the normal random error matrix with $\boldsymbol{\epsilon}_i \overset{i.i.d}{\sim} N(\boldsymbol{0}, \sigma^2\mathbb{I}_n)$. Here, $\mathbb{I}_n$ stands for a $n \times n$ diagonal matrix with all entries on the main diagonal equal to 1. In many applications, we would like to test contrasts: for feature i, the hypothesis is $\mathrm{H}_{0i} : \boldsymbol{q}^T\boldsymbol{\theta}_i = 0$. We can use the Wald's test to calculate P-values of $\mathrm{H}_{0,i}$. Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m)$ be the least squares estimator of $\boldsymbol{\theta}$. The Wald's statistic X_i^* and its corresponding P-value T_i is

$$X_i^* = \frac{(\boldsymbol{q}^T\hat{\boldsymbol{\theta}}_i)^2}{s^2\boldsymbol{q}^T(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{q}}, \quad T_i = 1 - F_0(X_i^*), \tag{2.4}$$

where $s^2 = \frac{1}{n-p_0}\left\|\boldsymbol{Y}_i - \boldsymbol{W}\hat{\boldsymbol{\theta}}_i\right\|_2^2$, and F_0 is the cumulative distribution function (CDF) of the $\chi^2(1)$ distribution. Here, T_is are not oracle P-values, but they are asymptotic oracle P-values. Details are provided in Lemma 2.1 and its proof in the Appendix A.

2.2.2 Two stages of DART

DART has two stages. In stage I, we transform the feature distance matrix into an aggregation tree where closer features are prioritized to be aggregated. In stage II, we embed multiple testing in the constructed aggregation tree and control the feature-level FDR.

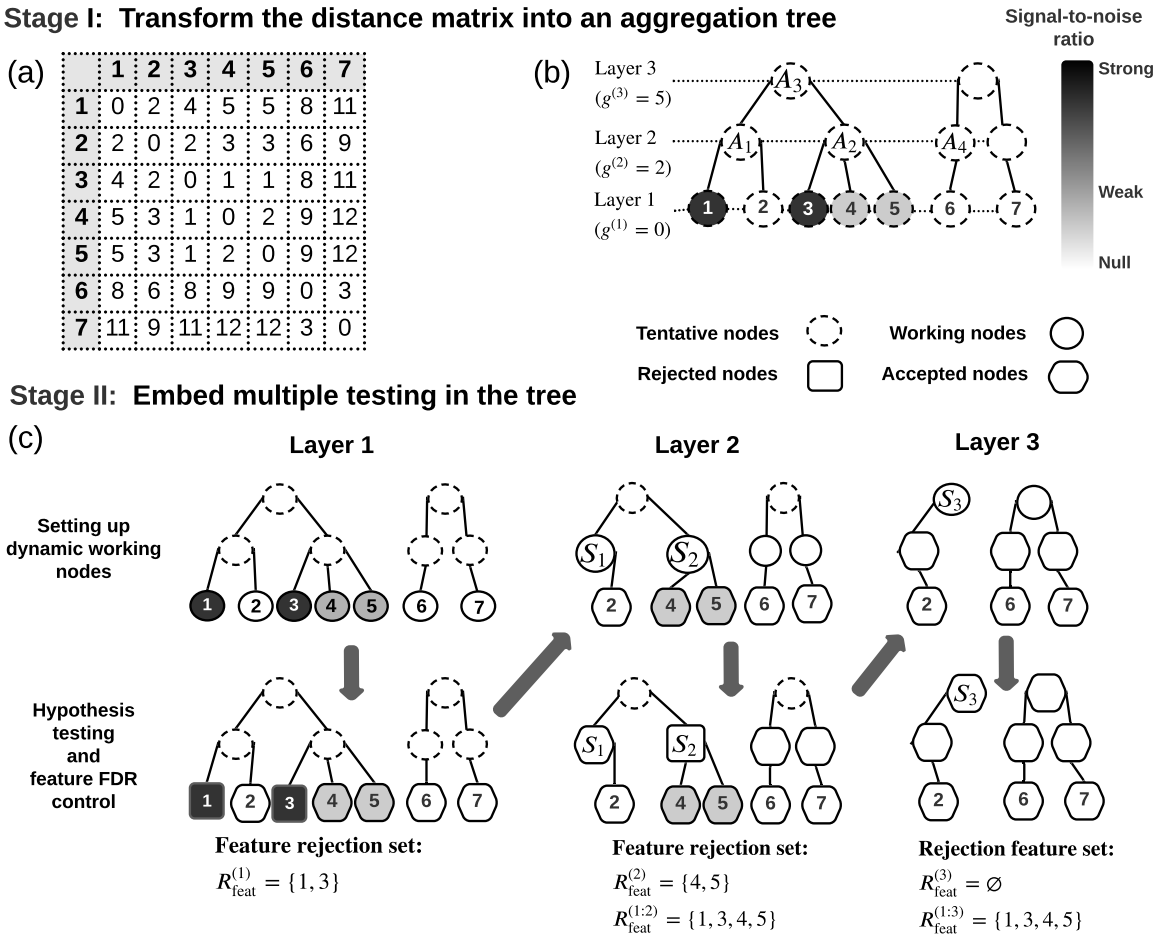
Stage I: Transform the distance matrix into an aggregation tree

Before we introduce the tree construction algorithm, we introduce some notations. Denote an L-layer aggregation tree by $\mathcal{T}_L = \{\mathcal{A}^{(\ell)} : \ell = 1, \ldots, L\}$, where $\mathcal{A}^{(\ell)}$ is the set of nodes on layer ℓ. Any node $A \in \mathcal{A}^{(\ell)}$ is a set of features. If a node A is aggregated from one or multiple children on layer $\ell - 1$, denote its children set by $\mathcal{C}(A)$. In other words, $A = \cup_{A' \in \mathcal{C}(A)} A'$; and $|\mathcal{C}(A)|$ counts the number of A's children. For example, in Figure 2.1b, $A_1 = \{1, 2\}$, $A_2 = \{3, 4, 5\}$, and $\mathcal{C}(A_3) = \{A_1, A_2\}$ with $|\mathcal{C}(A_3)| = 2$. A node could be equal to its child. For example, in Figure 2.1, $A_4 = \{6\}$ equal to its child. For any two nodes A and B (not necessarily on the same layer), the distance between A and B is $\text{dist}(A, B) = \max_{i \in A, j \in B} d_{ij}$.In Figure 2.1b, $\text{dist}(A_1, A_2) = 5$. The node distance defined here can be viewed as the complete linkage function initially proposed for hierarchical clusterings (Hastie et al., 2009). Under some special application context, other linkage functions may also be used. For any node A, the diameter of node A is $\text{dia}(A) = \max_{i \in A, j \in A} d_{ij}$. In Figure 2.1b, $\text{dia}(A_3) = 5$.

In stage I, we would like to construct an aggregation tree based on the feature distance matrix. On layer ℓ ($\ell \geq 2$), we hope that for all $A \in \mathcal{A}^{(\ell)}$

$$\text{dia}(A) \leq g^{(\ell)} \text{ and } |\mathcal{C}(A)| \leq M. \tag{2.5}$$

The threshold $g^{(\ell)}$ restricts the maximum distance among all features in the node



	1	2	3	4	5	6	7
1	0	2	4	5	5	8	11
2	2	0	2	3	3	6	9
3	4	2	0	1	1	8	11
4	5	3	1	0	2	9	12
5	5	3	1	2	0	9	12
6	8	6	8	9	9	0	3
7	11	9	11	12	12	3	0

FIGURE 2.1: An illustrating example of DART with 7 features. (a) Distance matrix of the 7 features. (b) In stage I, we transfer the distance matrix into the 3-layer aggregation tree based on Algorithm 3. The underlying feature signal-to-noise ratios are illustrated by the gray scales; these ratios are unknown. All nodes at this step are tentative. (c) In stage II, we perform the multiple testing embedded in the aggregation tree. We start from layer 1 and hierarchically proceed to higher layers. When testing on layer ℓ, all previous rejected features are excluded from the temporary nodes (dashed-line circled) to form the working nodes (solid-line circled) on this layer. The rejected nodes are marked by solid squares and the accepted nodes solid hexagons. All the features contained in the rejected nodes are rejected.

to make sure its containing features are close to each other; thus these features are likely to be co-null or co-alternative. If each of them has weak signal-to-noise ratios, aggregating them will boost their collective signal-to-noise ratio and increase their chance to be discovered. We restrict the nodes' children numbers to reduce the risk of creating mixed nodes (Definition 2.7) because too many mixed nodes will possibly lead to feature-level FDR inflation (see Section 2.3). To construct an aggregation tree satisfying 2.5, we propose an algorithm based on the Greedy algorithm (Cormen et al., 2001). The pseudo-code of this stage I algorithm is provided in Algorithm 3 in the the Appendix (Section A.1.1), along with its remarks.

At the end of stage I, an aggregation tree will be derived, with all nodes tentative. In stage II, based on the rejection path, we will further refine those nodes to form working nodes and working hypotheses.

Stage II: Embed multiple testing in the tree

The stage II testing procedure is recursive: On layer ℓ, the working hypotheses, the working P-values, and the P-value threshold depend on all previous layers.

On layer 1, leaf $\{i\}$ is coupled with the original hypothesis $\mathrm{H}_{0,i}$ in (2.1). We reject $\mathrm{H}_{0,i}$ if and only if the working P-value $T_i < \hat{t}^{(1)}(\alpha)$, where $\hat{t}^{(1)}(\alpha)$ is a threshold defined as follows.

$$\hat{t}^{(1)}(\alpha) = \left\{ \alpha_m \le t \le \alpha : \frac{mt}{\max\{\sum_{i=1}^{m} I(T_i < t), 1\}} \le \alpha, \right\} \tag{2.6}$$

where $\alpha_m = 1/\{m(\log m)^{1/2}\}$. This testing procedure is similar to the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995) with minor difference at the tail. Similar procedures have been proposed and discussed in other papers such as Liu et al. (2013) and Xie and Li (2018). After the testing procedure on layer 1, denote the rejected feature set by $R^{(1)}_{\text{feat}} = \{i : T_i \le \hat{t}^{(1)}(\alpha)\}$.

Traditional multiple testing procedure will stop on layer 1. However, DART will continue to aggregate nearby nodes because they are likely to be co-null or co-alternative. If the neighboring nodes all have weak signal-to-noise ratios, after aggregation their aggregated signal-to-noise ratio will be larger, and thus their chance to be discovered will increase..

On layer ℓ ($\ell \geq 2$), suppose the testing on the previous $\ell-1$ layers yields the rejected feature set $R_{\text{feat}}^{1:(\ell-1)} = \cup_{\ell'=1}^{\ell-1} R_{\text{feat}}^{(\ell')}$, where $R_{\text{feat}}^{(\ell')}$ is the rejected feature set on layer ℓ'. Denote the tentative node set on layer ℓ of the stage I aggregation tree by $\mathcal{A}^{(\ell)}$. For any tentative node $A \in \mathcal{A}^{(\ell)}$, we define $S(A) = A \setminus R_{\text{feat}}^{1:(\ell-1)}$. We call $S(A)$ a *working node*. The rejected features are removed from the working nodes because they have already been rejected and do not need to be tested again. For example, in Figure 2.1c, layer 1 rejected features 1 and 3; on layer 2, they are removed from the tentative nodes $A_1 = \{1, 2\}$ and $A_2 = \{3, 4, 5\}$ to form the working nodes $S_1 = \{2\}$ and $S_2 = \{4, 5\}$.

Define the testing node set on layer ℓ:

$$\mathcal{B}^{(\ell)} | \mathcal{Q}^{(1:\ell-1)} = \{S(A) : A \in \mathcal{A}^{(\ell)},\ |\mathcal{C}(S(A))| \geq 2\}.$$

Where $\mathcal{C}(S(A)) = \{A' \setminus R_{\text{feat}}^{1:(\ell-1)} : A' \in \mathcal{C}(A)\}$. We exclude the node with only one child because the node must have been tested on some lower layer. For example, in Figure 2.1c, $S_1 = \{2\}$ has been tested on layer 1. For any $S \in \mathcal{B}^{(\ell)}$, although it is dynamic, given the rejection path $\mathcal{Q}^{(1:\ell-1)}$, they are deterministic. Thus, conditioning on $\mathcal{Q}^{(1:\ell-1)}$, we construct the working node hypotheses:

$$\forall S \in \mathcal{B}^{(\ell)}, \quad \mathrm{H}_{0S} :\ \forall\ j \in S,\ j \in \Omega_0 \quad \text{versus} \quad \mathrm{H}_{1S} :\ \exists\ j \in S,\ j \in \Omega_1,$$

On layer ℓ, we aim to develop a multiple testing approach to simultaneously test these (conditional) working node hypotheses while asymptotically controls the feature-level FDR.

Definition 2.3 (Node working P-values)**.** For any node A, suppose T_j with $j \in A$ are the feature P-values. Then the node's working P-value is defined as

$$X_j = \bar{\Phi}^{-1}(T_j), \quad X_A = \sum_{j \in A} X_j / \sqrt{|A|}, \quad T_A = \bar{\Phi}(X_A), \tag{2.7}$$

where $\bar{\Phi}$ is the complementary CDF of the standard Gaussian distribution.

Noteworthy, working P-values are not oracle P-values. Because S is dynamic, the distribution of T_S depends on $\mathcal{Q}^{(1:\ell-1)}$. In Lemma 2.2 and 2.3, we will show T_S still has a good approximation to oracle P-value:

$$\sup_{S \in \mathcal{B}_0^{(\ell)}} \sup_{p \geq 1/m} \mathsf{P}(T_S \leq p | \mathcal{Q}^{(1:\ell-1)}) \leq p(1 + o(1)).$$

Similar to other multiple testing procedures, we threshold the working P-values to reject the nodes. For all $S \in \mathcal{B}^{(\ell)}$, H_{0S} is rejected if $T_S < \hat{t}^{(\ell)}(\alpha)$, where

$$\hat{t}^{(\ell)}(\alpha) = \sup \left\{ \alpha_m \leq t \leq \alpha : \frac{\sum_{\ell'=1}^{\ell-1} m^{(\ell')} \hat{t}^{(\ell')}(\alpha) + m^{(\ell)} t}{\max\{|R_{\text{feat}}^{(1:\ell-1)}| + \sum_{S \in \mathcal{B}^{(\ell)}} |S| I(T_S < t), 1\}} \leq \alpha \right\}, \tag{2.8}$$

Here $\alpha_m = 1/\{m(\log m)^{1/2}\}$ and $m^{(\ell)} = \sum_{S \in \mathcal{B}^{(\ell)}} |S|$. For simplicity sake, we use $\hat{t}^{(\ell)}$ to present $\hat{t}^{(\ell)}(\alpha)$ in the rest of the chapter. It is easy to see that $\hat{t}^{(\ell)}$ is recursive.

After applying the rejection rule (2.8), let

$$\mathcal{R}_{\text{node}}^{(\ell)} = \{S \in \mathcal{B}^{(\ell)} : \ T_S < \hat{t}^{(\ell)}\}, \quad R_{\text{feat}}^{(\ell)} = \cup_{S \in \mathcal{R}_{\text{node}}^{(\ell)}} S.$$

If a working node is rejected, We reject all its features. Although this rejection rule is aggressive, it is reasonable when most close features have co-null/co-alternative patterns. In Section 2.3, we will show this rule asymptotically controls feature-level FDR under mild conditions. The pseudo-code of the stage II algorithm is provided in Algorithm 4 in the Appendix A.1.2.

2.2.3 Tuning parameter selection

The number of total layers L, the maximum cardinality M, and the distance upper bounds $g^{(2)}, ..., g^{(L)}$ are viewed as tuning parameters. Here we provide a feasible approach to select the tuning parameters.

- $M = 3$. If M is too large, nodes on the aggregation tree are more likely to be mixed nodes (Definition 2.7) and the FDR will likely to be inflated. If M is too small, when weak signal-to-noise ratio features aggregate, their collective signal-to-noise ratios might still be too small to be identified. Numerical studies show that $M = 3$ performs well in practice.
- $L = \lceil \log_M m - \log_M c_m \rceil$, where c_m is the desired minimal number of working nodes on layer L. This is because on layer L, c_m will be lower bounded by m/M^L.
- The distance thresholds $g^{(1)}, \ldots, g^{(L)}$ are set recursively based on the criterion of maximizing the number of testable nodes on each layer. Let $g^{(1)} = 0$ and $G = \{g_1, \ldots, g_K\}$ be the candidate threshold set. On layer ℓ, let $G^{(\ell)} = \{g \in G : g > g^{(\ell-1)}\}$. For any $g \in G^{(\ell)}$, let $\mathcal{A}^{(\ell)}(g)$ be the resulting node set based on Algorithm 3. Then we set $g^{(\ell)}$ as

$$g^{(\ell)} = \arg\max_{g \in G^{(\ell)}} |\tilde{\mathcal{A}}^{(\ell)}(g)|, \quad \text{where } \tilde{\mathcal{A}}^{(\ell)}(g) = \{A : A \in \mathcal{A}^{(\ell)}(g),\ |\mathcal{C}(A)| \geq 2\}.$$

2.3 Asymptotic Theory

2.3.1 Weighted node-level FDR control

In multiple testing, type I error is commonly measured by the false discovery proportion (FDP) and its expectation, the false discovery rate (FDR). Under our model,

we defined the weighted node-level FDP and FDR up to layer ℓ as

$$\mathrm{FDP}_{\mathrm{node}}^{(1:\ell)} = \frac{\sum_{\ell'=1}^{\ell} \sum_{S \in \mathcal{R}_{\mathrm{node}}^{(\ell')} \cap \mathcal{B}_0^{(\ell')}} |S|}{\{\sum_{\ell'=1}^{\ell} \sum_{S \in \mathcal{R}_{\mathrm{node}}^{(\ell')}} |S|\} \vee 1}. \quad \mathrm{FDR}_{\mathrm{node}}^{(1:\ell)} = E(\mathrm{FDP}_{\mathrm{node}}^{(1:\ell)}),$$

Clearly, the denominator of $\mathrm{FDP}_{\mathrm{node}}^{(1:\ell)}$ counts the weighted number of all rejected nodes (taking maximum with 1 to avoid the denominator being 0), and numerator counts the weighted number of falsely rejected nodes; each node is weighted by its cardinality. Thus, a larger falsely rejected node will inflate the weighted node-level FDR more than a smaller falsely rejected node. We use the weight node-level FDP and FDR here because it can be more easily connected with the feature-level FDP and FDR. See Section 2.3.2.

To control the weighted node-level FDR, we introduce the following conditions.

Condition 2.1. Assume $m_1 \leq r_2 m^{r_1} \leq r_2 n^{r_1/r_3}$ for some $r_1 < (M^{L-1}+1)^{-1}$, $r_2 > 0$, and $r_3 > 0$.

Condition 2.1 assumes the important features are sparse, and the number of features is bounded by certain polynomial order of the sample size, $n \geq m^{r_3}$.

For any node A, we define its descendant set as

$$\mathcal{D}(A) = \{D : \exists\ \ell, \text{ such that } D \in \mathcal{A}^{(\ell)} \text{ and } D \subsetneqq A\}.$$

For example, in Figure 2.1b, $\mathcal{D}(A_3) = \{A_1, A_2, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$.

Definition 2.4 (Moderately strong Signal-to-Noise Ratio (SNR) nodes)**.** A node A is called a moderately strong SNR node if

$$\mathsf{P}\{T_A < \alpha_m,\ \forall D \in \mathcal{D}(A),\ T_D \geq \bar{\Phi}(m^{r_1-1}\sqrt{\log m})\} \geq C_1 > 0, \tag{2.9}$$

where α_m is the P-value thresholds lower bound defined in (2.8).

In fact, (2.9) is related to the alternative feature SNR. To better illustrate the moderately strong SNR nodes, we provide an equivalent definition when the test statistics follow the Normal distribution.

Example 2.2 (Normal distribution example). Suppose for feature i, a test statistic $Z_i \sim \mathrm{N}(\tau_i, 1)$ can be derived. The hypotheses are

$$\mathrm{H}_{0i} : \tau_i = 0 \quad \text{versus} \quad \mathrm{H}_{1i} : \tau_i \neq 0.$$

The P-values are $T_i = 2\bar{\Phi}(|Z_i|)$.

Under Example 2.2, a node A satisfying equation (9) when

$$\forall i \in A, \quad |\tau_i| \in [\gamma_m/\sqrt{|A|}, \beta_m/\sqrt{|A|-1}].$$

where

$$\beta_m = \sqrt{2(1-r_1)\log m - 2\log\log m}, \quad \gamma_m = \sqrt{2\log m + \log\log\log m}. \tag{2.10}$$

Although both β_m and γ_m increase with m, the rate is slow. In practice, when the sample size n increases, τ_i will increase with n, often at the rate of $\sqrt{n}$. Compared with $\sqrt{n}$, both β_m and γ_m are relatively small.

For any moderately strong SNR node A, suppose $A \in \mathcal{A}^{(\ell)}$. We will prove that with a certain non-vanishing probability, none of A's descendants will be rejected on the previous layers but A will be rejected on layer ℓ. On the tree $\mathcal{T}_L$, denote the set of all moderately strong SNR nodes by $\mathcal{A}_{\mathrm{md}}$. Define $c_{\mathrm{md}} = \min_{\ell \in \{1,\ldots,L\}} |\mathcal{A}_{\mathrm{md}} \cap \mathcal{A}^{(\ell)}|$ as the minimal number of moderately strong SNR nodes across all layers.

Condition 2.2. For some constant $r_4 > 0$, $c_{\mathrm{md}} \geq r_4 \log m$.

A node on layer ℓ has at most $M^{\ell-1}$ features, thus level ℓ has at least $M^{-\ell+1}m_1$ alternative nodes. Because we allow $m_1 = O(m^{r_1})$ by Condition 2.1, the total number

of the alternative nodes (containing alternative features) is also allowed to reach $O(m^{r_1})$. Condition 2.2 only requires $c_{\text{md}} \geq r_4 \log m$ among them are moderately strong SNR node; therefore, this condition is very weak.

For any node A on the top layer of $\mathcal{T}_L$, define its dependent node set as

$$\Gamma_A = \big\{A' \in \mathcal{A}^{(L)} : \{T_i, i \in A \cup A'\} \text{ are dependent}\big\}. \tag{2.11}$$

We assume Γ_A is relatively small for most of the As. We allow the existence of a) self-dependent nodes whose features are dependent and b) hub nodes which are dependent with many other nodes, but these nodes cannot be too many.

Condition 2.3 (Few self-dependent and hub nodes)**.** Define $\mathcal{A}' = \{A \in \mathcal{A}^{(L)} : |\Gamma_A| \geq \delta_{2m} = o(\sqrt{c_{\text{md}}})\}$. Assume $|\mathcal{A}'| = o(c_{\text{md}})$.

Under these conditions, the weighted node-level FDP of DART will be under control and thus also for the weighted node-level FDR.

Theorem 2.1 (Weighted node-level FDP and FDR control)**.** *Under Conditions 2.1-2.3, at any pre-specified level $\alpha \in (0,1)$, DART satisfies the following two statements.*

(1) For any $\epsilon > 0$, $\lim_{m,n\to\infty} \mathsf{P}(FDP^{(1:\ell)}_{\text{node}} \leq \alpha+\epsilon) = 1$. Consequently, $\lim_{m,n\to\infty} FDR^{(1:\ell)}_{\text{node}} \leq \alpha$.

(2) Let $\tilde{\Omega}_0 = \{j : \tilde{T}_j$ follows $\text{Unif}(0,1)\}$, where $\tilde{T}_j$ is the oracle P-value of feature j. If

$$\lim_{m\to\infty} |\tilde{\Omega}_0|/m = 1, \tag{2.12}$$

then for all $\epsilon > 0$,

$$\lim_{m,n\to\infty} \mathsf{P}(|FDP^{(1:\ell)}_{\text{node}} - \alpha| \leq \epsilon) = 1, \quad \lim_{m,n\to\infty} FDR^{(1:\ell)}_{\text{node}} = \alpha.$$

2.3.2 Feature-level FDR control

Define the feature-level FDP and FDR up to layer ℓ as

$$\mathrm{FDP}_{\mathrm{feat}}^{(1:\ell)} = \frac{|R_{\mathrm{feat}}^{(1:\ell)} \cap \Omega_0|}{|R_{\mathrm{feat}}^{(1:\ell)}| \vee 1}, \quad \mathrm{FDR}_{\mathrm{feat}}^{(1:\ell)} = E(\mathrm{FDP}_{\mathrm{feat}}^{(1:\ell)}).$$

It is easy to see that

$$\sum_{\ell'=1}^{\ell} \sum_{S \in \mathcal{R}_{\mathrm{node}}^{(\ell')}} |S| = \left| \cup_{\ell'=1}^{\ell} \cup_{S \in \mathcal{R}_{\mathrm{node}}^{(\ell')}} S \right| = \left| R_{\mathrm{feat}}^{(1:l)} \right|$$

$$\sum_{\ell'=1}^{\ell} \sum_{S \in \mathcal{R}_{\mathrm{node}}^{(\ell')} \cap \mathcal{B}_0^{(\ell')}} |S| = \left| \cup_{\ell'=1}^{\ell} \cup_{S \in \mathcal{R}_{\mathrm{node}}^{(\ell')} \cap \mathcal{B}_0^{(\ell')}} S \right| \leq \left| R_{\mathrm{feat}}^{(1:\ell)} \cap \Omega_0 \right|.$$

Thus $\mathrm{FDP}_{\mathrm{node}}^{(1:\ell)} \leq \mathrm{FDP}_{\mathrm{feat}}^{(1:\ell)}$. Controlling $\mathrm{FDR}_{\mathrm{node}}^{(1:\ell)}$ is easier than controlling $\mathrm{FDR}_{\mathrm{feat}}^{(1:\ell)}$. The challenge in controlling $\mathrm{FDR}_{\mathrm{feat}}^{(1:\ell)}$ lies in the existence of those nodes containing both null and alternative features. If such nodes are rejected, they are counted as true rejections for node-level weighted FDR control, but the null features in these nodes are counted as false rejections for feature-level FDR control.

Before we formally define those challenging nodes, we first define the strong SNR feature set $\Omega_{\mathrm{st}}^{(1:L)}$ and the weak SNR feature set Ω_{wk}.

Definition 2.5 (Strong SNR feature set)**.** Let $\Omega_{\mathrm{st}}^{(1:0)} = \emptyset$. On layer ℓ, recursively define

$$\mathcal{A}^{*,(\ell)} = \{A \setminus \Omega_{\mathrm{st}}^{(1:\ell-1)} : A \in \mathcal{A}^{(\ell)}\},$$

and the strong SNR node set as

$$\mathcal{G}_{\mathrm{st}}^{(\ell)} = \left\{ S \in \mathcal{A}^{*,(\ell)} : \ \forall j \in S, \ \mathsf{P}\{T_j \in \kappa(|S|)\} > 1 - o(m^{-r_1}) \right\}, \tag{2.13}$$

where $\kappa(S) = [m^{-\frac{1-r_1}{|S|-1}}, \left\{m(\log m \log\log m)^{1/2}\right\}^{-1/|S|}]$. Then the strong SNR feature

set on layer ℓ and up to layer ℓ are

$$\Omega_{\text{st}}^{(\ell)} = \cup_{S \in \mathcal{G}_{\text{st}}^{(\ell)}} S, \quad \Omega_{\text{st}}^{(1:\ell)} = \cup_{\ell'=1}^{\ell} \Omega_{\text{st}}^{(\ell')}.$$

Under Example 2.2, $\mathsf{P}\{T_j \in \kappa(|S|)\} > 1 - o(m^{-r_1})$ in (2.13) is satisfied when

$$|\tau_j| \in \left(\frac{\gamma_m}{\sqrt{|S|}} + \lambda_m, \frac{\beta_m}{\sqrt{|S|-1}} - \lambda_m \right),$$

where $\lambda_m = \sqrt{2r_1 \log m}$.

We will prove that with a probability converging to 1, none of the features in $\cup_{S \in \mathcal{G}_{\text{st}}^{(\ell)}} S$ will be rejected from layer 1 to layer $\ell - 1$ but all of them will be rejected on layer ℓ.

Definition 2.6 (Weak SNR feature set)**.** Let $\iota = (0, m^{\frac{r_1 - 1}{M^{L-1}}})$. Define the weak SNR feature set

$$\Omega_{\text{wk}} = \left\{ j \in \Omega_1 : \mathsf{P}(T_j \in \iota) = o(m^{-r_1}) \right\}. \tag{2.14}$$

Under Example 2.2, $\mathsf{P}(T_j \in \iota) = o(m^{-r_1})$ in (2.14) is satisfied when

$$|\tau_j| \in (0, \beta_m / \sqrt{M^{L-1}}).$$

When a node S contains only null features and weak signal features, then the probability of rejecting S is negligible.

Definition 2.7 (mixed nodes)**.** For any node $A \in \mathcal{A}^{(\ell)}$, let

$$A^* = A \setminus (\Omega_{\text{st}}^{(1:\ell-1)} \cup \Omega_{\text{wk}}), \quad A_0^* = A^* \cap \Omega_0, \qquad A_1^* = A^* \cap \Omega_1. \tag{2.15}$$

If $A_0^* \neq \emptyset$ and $A_1^* \neq \emptyset$, we call A a mixed node.

Noteworthy, not all nodes containing both null and alternative features are called mixed nodes. For example, suppose node $A \in \mathcal{A}^{(\ell)}$ have three child nodes $\mathcal{C}(A) =$

$\{A_1, A_2, A_3\}$, where $A_1 \subset \Omega_{\text{st}}^{(\ell-1)}$, $A_2 \subset \Omega_{\text{wk}}$, and A_3 contains all null features. Although both null and alternative features exist in node A, this is not a mixed node. This is because A_1 will be rejected on layer $\ell - 1$ with probability converging to 1 so that A's corresponding working node S is probably $A_2 \cup A_3$; also, S will not be rejected on layer ℓ with probability converging to 1 so that FDR will not be inflated. Define the strong and weak feature set will further narrow down the mixed nodes so that the condition to restrict their number (Condition 2.4) becomes weaker.

Condition 2.4 (Sparse mixed nodes)**.** Let $\mathcal{P} = \{S \in \mathcal{A}^{(L)} : S \text{ is a mixed node}\}$. Then $|\mathcal{P}| = o(c_{\text{md}})$.

Condition 2.4 assumes that the mixed nodes on layer L (the top layer) are rare. Equivalently, this means the dominating majority of the nodes contain: 1) only null features; 2) only alternative features; 3) a combination of null and alternative features but all alternative features are either weak or strong SNR features. Because the aggregation tree is constructed based on the distance matrix, this condition can be translated as how distance informs hypothesis states (null or alternative). To prove the consistence of the overall FDP, we need this condition because our rejection rule aggressively rejects all features in a node if the node is rejected. Without Condition 2.4 we might reject too many mixed nodes so that the feature-level FDR could be inflated.

Theorem 2.2 (Overall feature FDR control)**.** *Under Conditions 2.1-2.4, at any pre-specified level $\alpha \in (0, 1)$, DART satisfies the following two statements.*

(1) For any $\epsilon > 0$, $\lim_{m,n\to\infty} \mathsf{P}(FDP_{\text{feat}}^{(1:\ell)} \leq \alpha+\epsilon) = 1$. Consequently, $\lim_{m,n\to\infty} FDR_{\text{feat}}^{(1:\ell)} \leq \alpha$.

(2) If (2.12) holds, then for any $\epsilon > 0$,

$$\lim_{m,n\to\infty} \mathsf{P}(|FDP_{\text{feat}}^{(1:\ell)} - \alpha| \leq \epsilon) = 1, \quad \lim_{m,n\to\infty} FDR_{\text{feat}}^{(1:\ell)} = \alpha.$$

2.4 Numerical results

In this section, the simulation results are carried out to evaluate the performance of DART. We simulate m features located in the two-dimensional Euclidean space with randomly generated location coordinates: the first coordinate follows $\mathrm{N}(0, 2)$, and the second coordinate follows $\mathrm{Unif}(0, 4)$. A distance matrix $\boldsymbol{D} = (d_{i,j})_{m\times m}$ is calculated based on the feature location coordinates. Two different feature settings are considered, $m = 100$ and $m = 1000$. When $m = 100$, we generate $m_1 = 22$ alternative, and $n = 90$ samples. When $m = 1000$, we generate $m_1 = 141$ alternatives, and $n = 300$ samples.

Based on the tuning parameter selection criterion in Section 2.3, we construct a 2-layer aggregation tree when $m = 100$ and a 4-layers aggregation tree when $m = 1000$. More details about the tuning parameters settings and their selection procedure are shown in section 2.2.3.

We consider five different model settings, SE1–SE5. Throughout the five settings, the hypotheses are:

$$H_{0,i} : \theta_i = 0 \quad \text{against} \quad H_{1,i} : \theta_i \neq 0, \quad i \in \Omega.$$

SE1 simulates the working P-values satisfying the oracle P-value property, and thus mimics the ideal situation. SE2 and SE3 simulates the working P-values by misspecifying the null distributions, and thus these P-values do not satisfy the oracle P-value property. We use these two settings to evaluate the robustness of DART and the competing methods. SE4 simulates the linear regression model and SE5 simulates the Cox proportional hazard model. The feature P-values are derived from the Wald tests. We are interested to see how DART compares to the competing methods under these two commonly used models. Details in how to generate these simulation settings are displayed below (Section 2.4.1). Under each setting, the simulation is

repeated 200 times. The R codes are available at `https://github.com/xxli8080/DART_Code`.

2.4.1 Simulation Settings

Before we display the five settings, we first introduce the following notations that are used across all five settings:

$$\eta_{1,i} = \{[2\phi_1(d_{22,i}) - 0.2] \vee 0\} + \{\phi_2(d_{7,i})\};$$

$$\eta_{2,i} = \{[3.4\phi_3(d_{156,i}) - 0.8] \vee 0\} + 3\{\phi_4(d_{7,i})\}$$

$$+ 10 * I(i \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}),$$

where ϕ_1, ϕ_2, ϕ_3 and ϕ_4 are the PDF of $\mathrm{N}(0,1)$, $\mathrm{N}(0,0.1)$,$\mathrm{N}(0,0.8)$ and $\mathrm{N}(0,0.05)$, respectively.

SE1: For node $i \in \{1, ..., m\}$, the feature P-value $T_i = 2\bar{\Phi}(|\breve{Z}_i|)$, where the $\breve{Z}_1, \ldots, \breve{Z}_m$ are independently generated from $\mathrm{N}(\sqrt{n}\theta_i, 1)$, with

$$\theta_i = \begin{cases} \frac{1}{2}\eta_{1,i} I(\eta_{1,i} - 0.15 > 0), & (n,m) = (90,100) \\ \frac{1}{7}\eta_{2,i} I(\eta_{2,i} - 0.15 > 0), & (n,m) = (300,1000). \end{cases}$$

SE2: For node $i \in \{1, ..., m\}$, the feature P-value $T_i = 2\bar{\Phi}(|\breve{Z}_i|)$, where the $\breve{Z}_1, \ldots, \breve{Z}_m$ are independently generated from a mixture distribution $0.04\mathrm{Laplace}(\sqrt{n}\theta_i, 1)+$ $0.96\mathrm{N}(\sqrt{n}\theta_i, 1)$ with

$$\theta_i = \begin{cases} \frac{2}{5}\eta_{1,i} I(\eta_{1,i} - 0.15 > 0), & (n,m) = (90,100) \\ \frac{2}{13}\eta_{2,i} I(\eta_{2,i} - 0.15 > 0), & (n,m) = (300,1000) \end{cases}$$

SE3: For node $i \in \{1, ..., m\}$, the feature P-value $T_i = 2\bar{\Phi}(|\breve{Z}_i|)$, where the $\breve{Z}_1, \ldots, \breve{Z}_m$ are independently generated from the mixture distribution with $0.04\mathrm{t}_5(\sqrt{n}\theta_i)+$ $0.95\mathrm{N}(\sqrt{n}\theta_i, 1)$. Here, $\mathrm{t}_5(\sqrt{n}\theta_i)$ stands for the student t distribution with 5 degree of freedom and none centrality parameter $\sqrt{n}\theta_i$, with

$$\theta_i = \begin{cases} \frac{1}{3}\eta_{1,i} I(\eta_{1,i} - 0.15 > 0), & (n,m) = (90,100) \\ \frac{2}{13}\eta_{2,i} I(\eta_{2,i} - 0.15 > 0), & (n,m) = (300,1000) \end{cases}$$

SE4: Consider the linear mode defined in (2.3), with $p_0 = 3$, and $\sigma = 1$. In model (2.3), $W_1 = 1$ is the intercept term, W_2 and W_3 are sampled from Binom(0.5) and Unif$(0.1, 0.5)$, respectively. Also let $\theta_{1,i} = \theta_{3,i} = 0.1$ and

$$\theta_i = \theta_{1,i} = \begin{cases} 2\eta_{1,i} I(\eta_{1,i} - 0.15 > 0), & (n, m) = (90, 100) \\ \frac{5}{6}\eta_{2,i} I(\eta_{2,i} - 0.15 > 0), & (n, m) = (300, 1000) \end{cases}$$

The feature P-value T_i is defined in (2.4).

SE5: Consider the Cox regression model

$$\lambda_i(t) = \lambda_{0i}(t) \exp\{\theta_{1,i} W_1 + \theta_{2,i} W_2\}$$

Where $\lambda_i(t)$ and $\lambda_{0i}(t)$ is the hazard and baseline hazard at time t, respectively. Set $\theta_{0,i} = \theta_{2,i} = 0.1$ and

$$\theta_i = \theta_{1,i} = \begin{cases} \frac{4}{5}\eta_{1,i} I(\eta_{1,i} - 0.15 > 0), & (n, m) = (90, 100) \\ \frac{2}{7}\eta_{2,i} I(\eta_{2,i} - 0.15 > 0), & (n, m) = (300, 1000) \end{cases}$$

The covariates W_1 and W_2 are sampled from Binom(0.5) and Unif$(0.1, 0.5)$, respectively. The event time is generated from the exponential distribution with rate $\exp\{\theta_{1,i} W_1 + \theta_{2,i} W_2\}$, and the censoring time is sampled from Unif$(0, 5)$. The feature P-value T_i is obtained from the Wald test.

We set the nominal FDR at the level $\alpha = 0.05, 0.1, 0.15, 0.20$. We follow Section 2.2.3 to select the tuning parameters; details are displayed in the Appendix A.1.3.

2.4.2 Numerical Results

DART successfully controlled the empirical FDR under the desired level. The FDR control is robust when the model is misspecified. Figure 2.2 shows how DART performs when the algorithm stopped at different layers. Obviously, the one-layer DART is the same as the traditional single layer multiple testing method which ignores the distance matrix. As the number of maximum layers L goes up, more

alternative features are aggregated and identified. Notably, increasing the nominal FDR level cannot lead to such great increase in sensitivity.

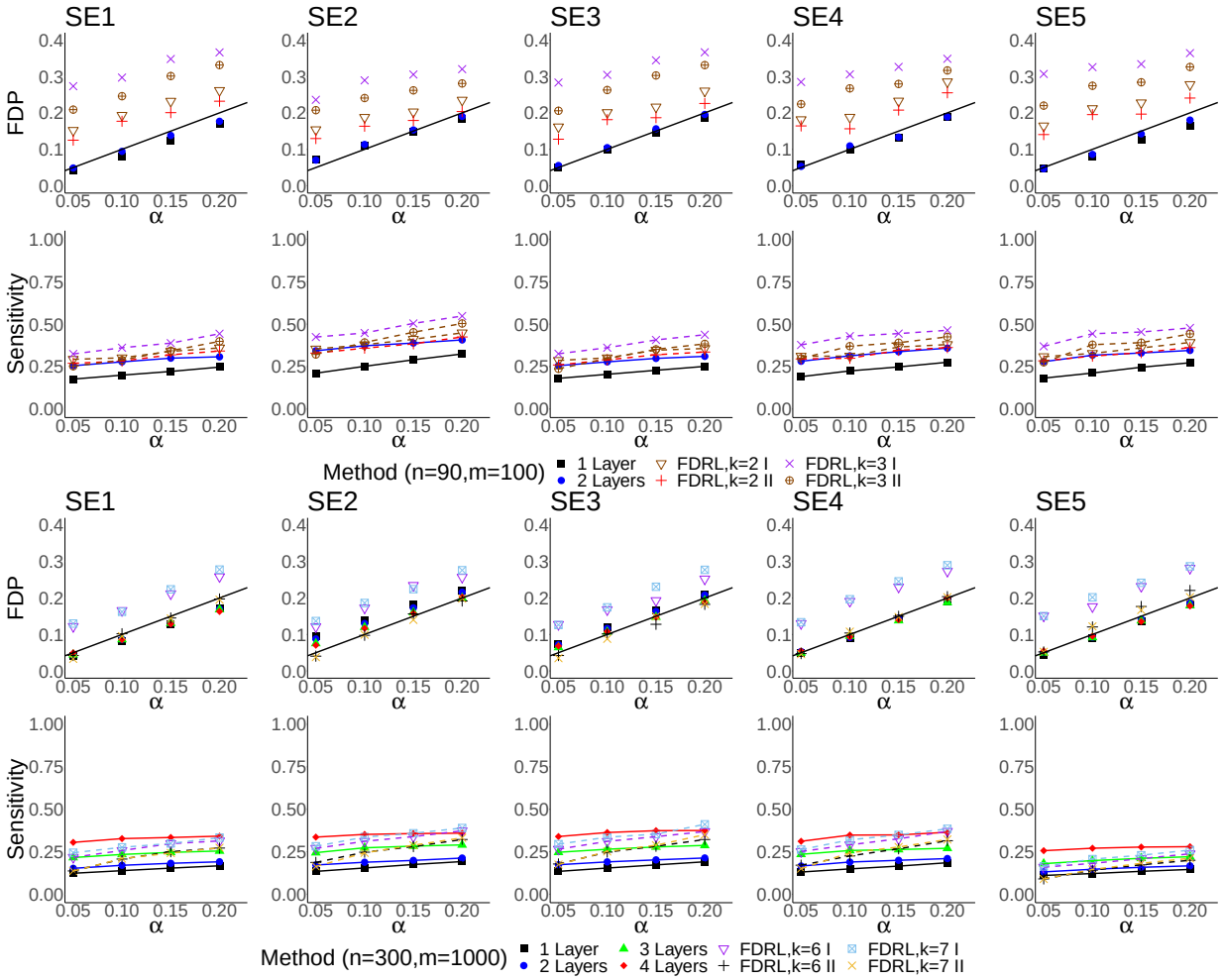


FIGURE 2.2: Simulation results for setting SE1-SE5. The first two rows represent the results in the setting $(n, m) = (90, 100)$, and the second two rows represent the results in the setting $(n, m) = (300, 1000)$.

We also compared the performance of DART with the two FDR_L procedures (FDR_L I and FDR_L II) proposed by Zhang et al. (2011). The two procedure adjust each feature's P-value according to its k-nearest neighbors; the adjusted P-value is the median of its neighborhood P-values. The FDR_L I and FDR_L II procedures use different methods to estimate the distribution of the adjusted P-values. We compared to both procedures in our simulation. To perform a fair comparison, we also tried

a wide range of choices $k \in \{2, 3, \ldots, 9\}$. When $(n, m) = (90, 100)$, both FDR_L I and FDR_L II led to FDR inflation regardless of the choice of k. When $(n, m) = (300, 1000)$, the FDR_L I procedure constantly led to inflated FDR, while the FDR_L II procedure led to the desired FDR when $k \leq 7$ but the sensitivity is lower than DART. Figure 2.2 presents the performance of FDR_L procedures with $k = 2, 3$ when $(n, m) = (90, 100)$ and $k = 6, 7$ when $(n, m) = (300, 1000)$, because under these k settings, the FDR_L procedures perform the best.

One reason that the FDR_L procedures did not perform as well as DART is that the methods use a constant k to aggregate P-values. Very often, the distance among features often cannot be fully captured by the neighborhood with constant number of neighbors. For example, an feature far away from all other features also have k nearest neighbors; however, the isolated feature and its neighbors often do not share co-importance. Thus, FDR_L does not perform well under these settings.

2.5 Data Analysis

We apply DART to a clinical trial on the hematopoietic stem cell transplantation (HCT), where microbiome data are collected from 144 leukemia patients before and after the HCT. Graft-versus-host disease (GVHD) is one of the major complications of the HCT. Recent studies have linked GVHD to the disruptions of the gut microbiome (Jenq et al., 2012), and the disruptions may be related to the environmental changes such as post-transplant care (Claesson et al., 2012). The goal of this study is to investigate the potential impact of the post-transplant care (home care versus standard hospital care) on the patient gut microbiota composition.

To achieve the goal, the patient fecal samples are collected before and after HCT; the fecal microbiome are sequenced by the 16S ribosomal RNA sequencing at the Memorial Sloan Kettering Cancer Center. The data are then pre-processed by the R pack-

age, `DADA2` (Callahan et al., 2016), to generate the amplicon sequence variants (ASV) and the read counts. Samples with less than 200 total read counts and the ASVs with read counts fewer than 4 in more than 80% of the samples are removed from the analysis. After the pre-processing procedure, the data set contains 288 samples (before- and after- HCT) from 144 patients, each with 97 ASVs. The data are available at `https://github.com/xxli8080/DART_Code/tree/master/Data_Analysis`. In our analysis, to increase computation stability, the zero counts are replaced by 0.5 (Aitchison, 1982; Kurtz et al., 2015).

In microbiome studies, the ASV abundance compositions are more meaningful than the absolute read counts. To modeling the compositional microbiome data, we use the additive log-ratio transformation proposed by Aitchison (1982). Specifically, we choose the most abundant ASV (the ASV with the largest median read counts across all patients) as the reference ASV, and define M_i as the log read counts ratio between the ASV i and the reference ASV. For example, for a patient, if the read counts of ASV i and the reference ASV are 100 and 200 respectively, then $M_i = \log(100/200) = -\log 2$.

Because one ASV is chosen as the reference ASV, the distance matrix is calculated among the remaining 96 non-reference ASV using the R package `Phangorn` (Schliep, 2011) based on the JC69 model (Jukes et al., 1969). The JC69 model is a classical Markov model of DNA sequence evolution and can be used to estimate the evolutionary distance between sequences. Two ASVs with similar sequences tend to be close with each other, and more likely to perform similar biological functions. Therefore, we will incorporate the distance matrix in identifying the important ASVs. We use the linear model, defined in (2.3) with $p_0 = 3$, to regress the microbiome composition changes before and after HCT on the after-transplantation care (home care vs. hospital care) and other covariates. Specifically, for the non-reference ASV i,

$i \in \{1, \ldots, 96\}$,

$$M_{1,i} - M_{0,i} = \theta_{1,i} W_1 + \theta_{2,i} W_2 + \theta_{3,i} W_3 + \epsilon_i \tag{2.16}$$

Here, $M_{0,i}$ (and $M_{1,i}$) is the log counts ratio between ASV i and the reference ASV before (and after) the transplant. Thus $M_{1,i} - M_{0,i}$ is the corresponding Y_i in the model (2.3). In addition, $W_1 = 1$ is the intercept term, W_2 is the type of care, W_3 is the length of the care (the gap between the HCT surgery and the after-care sample collection), and the $\epsilon \sim N(0, \sigma^2)$ is the random error term with unknown σ^2. To check whether the after-transplant care affects the ASV compositions, we set up the hypotheses $H_{0i} : \theta_{2,i} = 0$, $i = 1, \ldots, 96$. The P-value T_i are calculated based on the Wald tests.

Based on the tuning parameter selection procedure described in Section 2.3, we construct an aggregation tree with $M = 3$, $L = \lceil \log_M 96 - \log_M 30 \rceil = 2$, and $g^{(2)} = 8/\sqrt{144 \log 96 \log\log 96}$. The aggregation tree has 33 non-single-child nodes on the second layer. The nominal FDR level is set at 0.1.

The performance of the DART is compared with two competing methods: 1) BH procedure; 2) FDR$_L$. For the FDR$_L$ I and II procedures, we considered $k = 2$ or 3. Figure 2.3(a) shows that the ASVs that are close to each other tend to have similar (small or large) P-values. This suggests that the co-importance pattern among similar ASVs might hold here. In the end, the two-layer DART identified 9 important ASVs while the traditional BH procedure did not identify any ASV. Both FDR$_L$ I and FDR$_L$ II procedures identified 14 important ASVs when $k = 2$. When $k = 3$, FDR$_L$ I identified 16 important ASVs, and FDR$_L$ II identified 7 important ASVs.

In order to evaluate the stability of these methods, we conduct the bootstrap with 200 re-samplings. For a specific testing method, the rejection rate of an ASV is calculated as the ratio of the times that the ASV is identified in the 200 rounds of resamplings. If a method is stable, an ASV should tend to be consistently rejected or

accepted. In other words, for a valid and powerful test, most null ASVs are expected to have small rejection rates, and very few alternative ASVs are expected to have high rejection rates. Figure 2.3(b) shows that DART and BH procedures generates the histograms with a peak rejection rate within $[0, 0.2)$, while FDR_L have the peak rejection rate between $[0.1, 0.3)$. Table 2.1 listed the proportion of ASVs with large (> 0.8) or small (≤ 0.1) rejection rates for each method. Compared with FDR_L method, DART and BH have a higher proportion of ASVs with small rejection rates, indicating both DART and BH have lower risk in FDR inflation. Meanwhile, FDR_L methods have a small proportions of ASVs with the rejection rate within $0 - 0.1$, indicating it is not stable in accepting null ASVs. On the other hand, DART also has a higher proportion of ASVs with large rejection rates comparing to the BH method. This indicates that DART has a robust high power.

Table 2.1: Summary of the boostrap results. (RR stands for rejection rates)

Method	RR≤ 0.1	RR> 0.8
DART	0.27	0.02
BH	0.47	0
FDR_L I, $k = 2$	0.01	0.03
FDR_L I, $k = 3$	0	0.04
FDR_L II, $k = 2$	0.09	0.02
FDR_L II, $k = 3$	0.1	0.03

2.6 Discussion

In this chapter, we develop a novel multiple testing method, DART, to incorporate feature distance in multiple testing. Under many application contexts, the feature distances serve as auxiliary information of their co-importance pattern. DART utilizes this information to boost the testing power. DART applies to the P-values obtained from many asymptotic tests, and thus can work with a wide range of models.

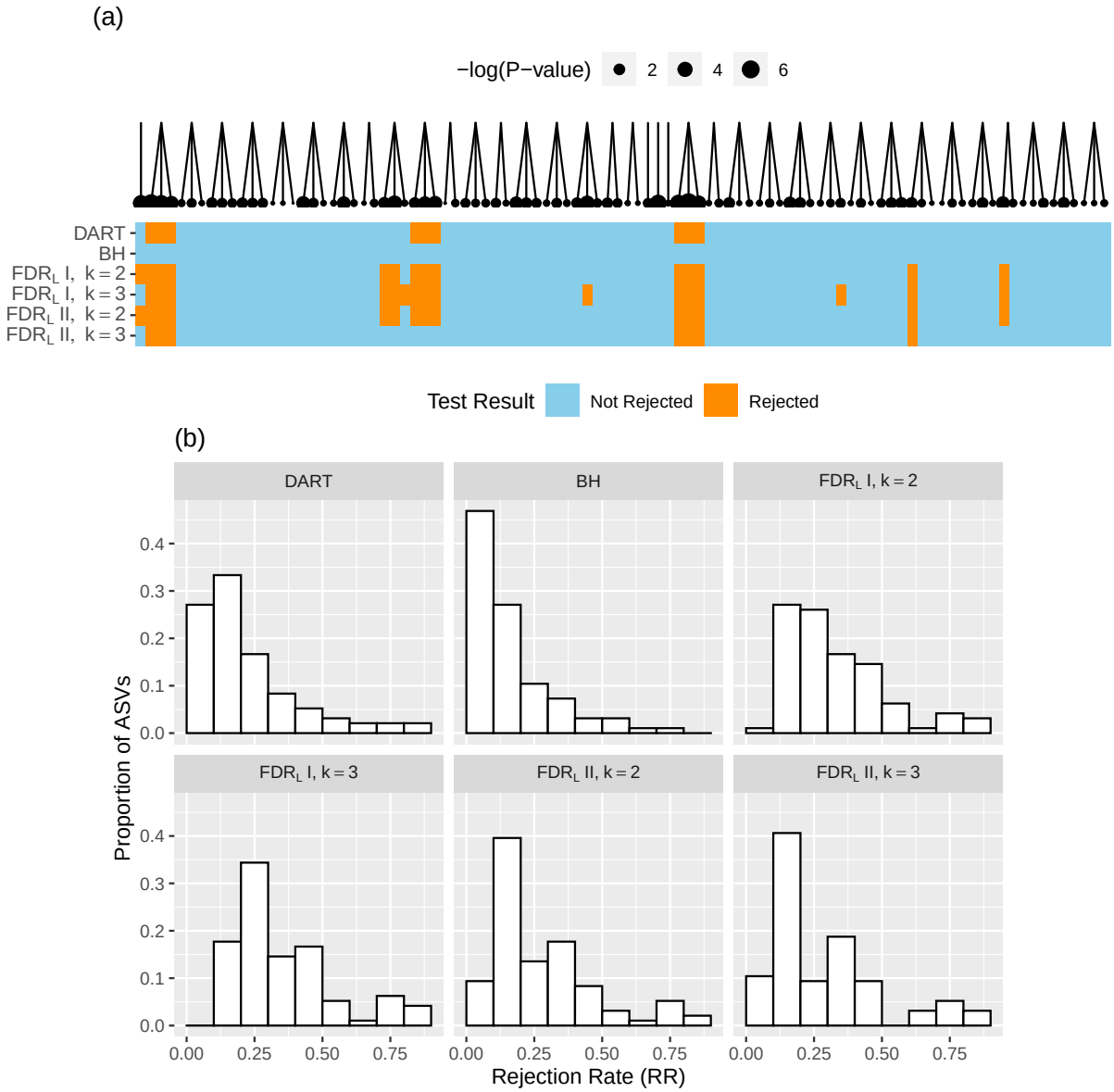


FIGURE 2.3: (a) Illustration of the leaf P-values, the aggregation tree, and the testing results. The leaf size is scaled according to the inverse of the P-values. The testing results of different methods are shown in different rows, with blue blocks representing the accepted non-reference ASVs and orange blocks representing the rejected non-reference ASVs. (b) Histograms of ASV rejection rate across the bootstrap with 200 re-samplings.

Stage 1 of DART involves constructing an aggregation tree. We provided Algorithm 3 to construct the aggregation tree. Other algorithms may also work, and result in a different aggregation tree from the same distance matrix. Consequently, Stage 2 testing process could lead to different results based on different trees. In practice, if several aggregation trees exist, DART can be applied to all of them, and we can take the one with the most rejections. The asymptotic validity will still hold for this procedure.

DART is a multiple testing method embedded in a hierarchical tree that constructed from the distance matrix. It can be easily extended to the case where other information implies the co-importance pattern of the features. Such information could from domain knowledge, external data sets, or other resources. In addition, the hierarchical testing ideas and techniques can also be extended to solve other multiple testing problems.

2.7 Proofs of the Main Results

Before the proof, we need to introduce some further notations. On layer ℓ, for a working node $S \in \mathcal{B}^{(\ell)}$, let $\mathcal{U}(S) = \{S' \subset S : S' \in \cup_{\ell'=1}^{\ell-1} \mathcal{B}^{(\ell')}\}$ be the collection of sets in the testing path of S. In addition, let $\mathcal{U}^c(S) = \{S'' \in \cup_{\ell'=1}^{\ell-1} \mathcal{B}^{(\ell')} : S'' \cap S = \emptyset, S'' \cup S \subset A, \text{ for some } A \in \mathcal{A}^{(\ell)}\}$ be the collection of sets that was planning to combined with S on layer ℓ of the static aggregation tree but rejected on previous layers. When $S \in \mathcal{B}^{(1)}$, we set $\mathcal{U}(S) = \mathcal{U}^c(S) = \emptyset$. We define $G_S(c)$ as the complementary CDF conditional on previous testing results. When $\ell = 1$, we have $S = \{i\} \subset \{1, ..., m\}$, and $G_S(c) = P(Z_i \geq c)$ with $Z_1, \ldots, Z_m \overset{iid}{\sim} N(0, 1)$. When $\ell > 1$, the oracle rejection path for set $S \in \mathcal{B}^{(\ell)}$ is recursively defined as

$$\mathcal{Q}_z^{(1:\ell-1)} = \{z : \forall S' \in \mathcal{U}(S), G_{S'}(Z_{S'}) \geq \hat{t}^{(\ell_{S'})}(\alpha), \forall S'' \in \mathcal{U}^c(S), G_{S''}(Z_{S''}) \leq \hat{t}^{(\ell_{S''})}(\alpha)\},$$

where

$$G_S(c) = \mathsf{P}\big(Z_S \geq c \big| \mathcal{Q}_z^{(1:\ell-1)}\big)$$

and $Z_S = \sum_{i\in S} Z_i/\sqrt{|S|}$, and $\ell_{S'}, \ell_{S''} \in \{1, ..., \ell-1\}$ is the value s.t. $S' \in \mathcal{B}^{(\ell_{S'})}$ and $S'' \in \mathcal{B}^{(\ell_{S''})}$, respectively.

Given $Z_1, \ldots, Z_m$ are mutually independent, we have

$$G_S(c) = \mathsf{P}\big(Z_S \geq c \big| \forall S' \in \mathcal{U}(S), G_{S'}(Z_{S'}) \geq \hat{t}^{(\ell_{S'})}(\alpha)\big)$$

Given the definition of $G_S(c)$, we define the rejection path as

$$\mathcal{Q}^{(1:\ell-1)} = \{x : \forall S' \in \mathcal{U}(S), G_{S'}(X_{S'}) \geq \hat{t}^{(\ell_{S'})}(\alpha), \forall S'' \in \mathcal{U}^c(S), G_{S''}(X_{S''}) \leq \hat{t}^{(\ell_{S''})}(\alpha)\} \tag{2.17}$$

In addition, for two sequence of real numbers a_m and b_m, we write $a_m = o(b_m)$ when $a_m/b_m \to 0$, and $a_m = O(b_m)$ when $\lim_{m\to\infty} |a_m/b_m| \leq C$ for some constant C. To prove the asymptotic properties of DART, we need the following lemmas.

Lemma 2.1. *Under the linear regression model* (2.3)*, T_is are asymptotic oracle P-values.*

Lemma 2.2. *Let $\mathcal{P}_i = \{p \in [0,1] : \mathsf{P}(\tilde{T}_i < p) \geq \epsilon(m)\}$ and $\mathcal{P}_i' = \{p \in [0,1] : \mathsf{P}(\tilde{T}_i < p) \geq \epsilon(m)\epsilon'(m)\}$, with $\epsilon(m), \epsilon'(m) \to 0$. For any set of independent random variable $\hat{T}_i \in [0,1]$, and a collection $\mathcal{M} = \{S \subset \{1, ..., m\} : |S| < c_0\}$ with some constant c_0,*

(1) If $\max_{i\in\mathcal{M}} \sup_{p\in\mathcal{P}_i'} \big|\mathsf{P}(\hat{T}_i < p)/\mathsf{P}(\tilde{T}_i < p) - 1\big| \to 0$*, then,*

$$\sup_{S_0\in\mathcal{M}} \sup_{p\geq\epsilon(m)} \left|\frac{\mathsf{P}(\sum_{i\in S_0} \hat{X}_i > c_{S_0}(p))}{\mathsf{P}(\sum_{i\in S_0} \tilde{X}_i > c_{S_0}(p))} - 1\right| \to 0,$$

(2) If $\lim_{m\to\infty} \max_{i\in\mathcal{M}} \sup_{p\in\mathcal{P}_i'} \big(P(\hat{T}_i < p)/\mathsf{P}(\tilde{T}_i < p) - 1\big) \leq 0$*, then,*

$$\lim_{m\to\infty} \sup_{S_0\in\mathcal{M}} \sup_{p\geq\epsilon(m)} \left(\frac{\mathsf{P}(\sum_{i\in S_0} \hat{X}_i > c_{S_0}(p))}{\mathsf{P}(\sum_{i\in S_0} \tilde{X}_i > c_{S_0}(p))} - 1\right) \leq 0$$

Here, $\hat{X}_i = \bar{\Phi}^{-1}(\hat{T}_i)$, $\tilde{X}_i = \bar{\Phi}^{-1}(\tilde{T}_i)$ *and* $c_{S_0}(p)$ *is the value s.t.* $\mathsf{P}[\sum_{i\in S_0} \tilde{X}_i > c_{S_0}(p)] = p$.

Lemma 2.3. *Let* $\tilde{\Omega}_0 = \{i : \tilde{T}_i$ *follows* $\mathrm{Unif}(0,1)\}$, $\mathcal{B}_{0a}^{(\ell)} := \{S \in \mathcal{B}_0^{(\ell)} : \exists A \in \mathcal{A}^{(L)} \setminus \mathcal{A}', s.t. S \subset A\}$, *and* $\mathcal{B}_{0b}^{(\ell)} := \{S \in \mathcal{B}_{0a}^{(\ell)} : S \in \tilde{\Omega}_0\}$, *we have:*

(1) $$\max_{S\in\mathcal{B}_{0a}^{(\ell)}} \sup_{c\in[0,\gamma_m]} \left|\frac{G_S(c)}{\bar{\Phi}(c)} - 1\right| \to 0$$

(2) $$\max_{S\in\mathcal{B}_{0b}^{(\ell)}} \sup_{c\in[0,\bar{\Phi}^{-1}(1/m)]} \left|\frac{\mathsf{P}(X_S > c|\mathcal{Q}^{(1:\ell-1)})}{\mathsf{P}(X_S > c)} - 1\right| \to 0$$

Lemma 2.4. *Define*

$$\mathcal{X}^{(\ell)} = \left\{x : \sum_{S\in\mathcal{B}_0^{(\ell)}} |S| I(T_S < \hat{t}^{(\ell)}) - \sum_{S\in\mathcal{B}_0^{(\ell)}} |S|\hat{t}^{(\ell)} \le \left\{\sum_{S\in\mathcal{B}_0^{(\ell)}} |S|\hat{t}^{(\ell)}\right\}\epsilon\right\} \tag{2.18}$$

$$\mathcal{X}'^{(\ell)} = \left\{x : \left|\frac{\sum_{S\in\mathcal{B}_0^{(\ell)}} |S| I(T_S < \hat{t}^{(\ell)})}{\sum_{S\in\mathcal{B}_0^{(\ell)}} |S|\hat{t}^{(\ell)}} - 1\right| \ge \epsilon\right\}$$

Then, $\forall \ell = 1, ..., L$, *when the FDR control holds on layer* $1, ..., \ell - 1$,

(1) For all $\epsilon \in (0, \alpha)$, *if* $\mathsf{P}(m\hat{t}^{(\ell)} \ge Cc_{md}) \to 1$, *then* $\mathsf{P}(\mathcal{X}^{(\ell)}) = 1 - o(1)$. *Together with* $\lim_{m\to\infty} |\tilde{\Omega}_0|/m = 1$, *we have* $P(\mathcal{X}'^{(\ell)}) = 1 - o(1)$.

(2) On $\cap_{h=1}^{\ell} \mathcal{X}^{(h)}$, *there exist a constant* C *s.t.* $\hat{t}^{(\ell)} \le Cm^{r_1-1}$.

(3) Let $\hat{c}_S$ *be the rejection threshold for the test node* $S \in \mathcal{B}^{(\ell)}$, *s.t.* $\bar{G}_S(\hat{c}_S) = \hat{t}^{(\ell)}$. *Then on* $\cap_{h=1}^{\ell} \mathcal{X}^{(h)}$,

$$\hat{c}_S > \beta_m, \ \forall S \in \mathcal{B}^{(\ell)},$$

and on $\cap_{h=1}^{\ell-1} \mathcal{X}^{(h)}$,

$$\hat{c}_S < \gamma_m, \ \forall S \in \mathcal{B}^{(\ell)}.$$

Lemma 2.5.

$$\frac{\sum_{S\in\mathcal{B}_0^{(\ell)}} |S|\hat{t}^{(\ell)}}{\sum_{S\in\mathcal{B}^{(\ell)}} |S| I(T_S < \hat{t}^{(\ell)})} = \alpha(1 + o(1)) \tag{2.19}$$

Proof of Theorem 2.1. Since the proof of the theorem statement (2) is similar to the proof of the theorem statement (1), we will only focusing on the proof of statement (1).

The random variable $FDP^{(\ell)}$ can be decomposed to the product of two parts.

$$FDP^{(\ell)} = \frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|I\{T_S < \hat{t}^{(\ell)}\}}{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}} \times \frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}}{\max(\sum_{S\in\mathcal{B}^{(\ell)}}|S|I\{T_S < \hat{t}^{(\ell)}\}, 1)} \tag{2.20}$$

Based on (2.20), in order to prove $\lim_{m\to\infty}\mathsf{P}(FDP^{(\ell)} \le \alpha+\epsilon) = 1$ for all $\epsilon > 0$, we only need prove

$$\lim_{m\to\infty}\mathsf{P}\left\{\frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|I\{T_S < \hat{t}^{(\ell)}\}}{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}} - 1 < \epsilon\right\} \to 1 \tag{2.21}$$

$$\lim_{m\to\infty}\mathsf{P}\left\{\left|\frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}}{\max(\sum_{S\in\mathcal{B}^{(\ell)}}|S|I\{T_S < \hat{t}^{(\ell)}\}, 1)} - \alpha\right| > \epsilon\right\} \to 0 \tag{2.22}$$

(2.22) is immediately followed by Lemma 2.5, and we will prove (2.21) by induction. Below is a list of the proof sketch:

1. On layer 1, show $P(m\hat{t}^{(1)} \ge Cc_{\mathrm{md}}) \to 1$. Then, by applying Lemma 2.4, we have
   - $P(\mathcal{X}^{(1)}) \to 1$, which is equivalent to (2.21). Hence, we proved the FDR control on layer 1.
   - $P(\beta_m < \hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(1)}) \to 1$, and $P(\hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(2)}) \to 1$. Note that although this conclusion is not used to prove the FDR control on the current layer, but is necessary to guarantee the FDR control on higher layers.
2. On layer $\ell \ge 2$, assume the FDR control holds on previous layers and $P(\mathcal{X}^{(\ell')}) \to 1$ for all $\ell' = 1, \ldots, \ell-1$. Then by Lemma 2.4, $P(\beta_m < \hat{c}_S < \gamma_m, \forall S \in \cup_{\ell'=1}^{\ell-1}\mathcal{B}^{(\ell')}) \to 1$, and $P(\hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(\ell)}) \to 1$. Accordingly, we can get $P(m\hat{t}^{(1)} \ge Cc_{\mathrm{md}}) \to 1$. Then, by applying the Lemma 2.4 again, we have

- $P(\mathcal{X}^{(\ell)}) \to 1$, which is equivalent to (2.21). Hence, we proved the FDR control on layer ℓ.
- $P(\beta_m < \hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(\ell)}) \to 1$, and $P(\hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(\ell+1)}) \to 1$.

We start the proof on layer 1.

Layer 1:

Take a subset $\mathcal{F}^{(1)} \subset \mathcal{A}_{\text{md}} \cap \mathcal{A}^{(1)}$, such that $|\mathcal{F}^{(1)}| = c_{\text{md}}$. For any $i \in \mathcal{F}^{(1)}$, we have $\mathsf{P}(X_i > \gamma_m) \geq C$. By Markov's inequality, we have:

$$\mathsf{P}\big(\big|\sum_{i\in\mathcal{F}^{(1)}} I(X_i > \gamma_m) - \sum_{i\in\mathcal{F}^{(1)}} \mathsf{P}(X_i > \gamma_m)\big| \geq c_{\text{md}}^{3/4}\big) \leq C(c_{\text{md}})^{-1/2}$$

Thus,

$$\mathsf{P}\big[\sum_{1\leq i\leq m} I(T_i \leq \hat{t}^{(1)}) \geq Cc_{\text{md}} - c_{\text{md}}^{3/4}\big] \geq 1 - o(1)$$

Therefore, by Lemma 2.5, exists constant $C^{(1)}$, s.t.

$$\mathsf{P}\big[m_0\hat{t}^{(1)} \geq C^{(1)}c_{\text{md}}\big] \geq 1 - o(1) \tag{2.23}$$

Together with Lemma 2.4 (1), we have $\mathsf{P}(\mathcal{X}^{(1)}) \to 1$ and accordingly, $\mathsf{P}(FDP^{(1)} < \alpha + \epsilon) \to 1$.

Layer ℓ:

Based on similar arguments on Layer 1, it is suffice to show $\mathsf{P}(m_0\hat{t}^{(\ell)} > C^{(\ell)}c_{\text{md}}) \to 1$ for some constant $C^{(\ell)}$.

Assume $\forall h = 1, \ldots, \ell - 1$, $\mathsf{P}(\mathcal{X}^{(h)}) \to 1$, then by Lemma 2.4, we have $\mathsf{P}(\beta_m < \hat{c}_S < \gamma_m, \forall S \in \mathcal{B}^{(h)}) \to 1$, and $\mathsf{P}(c_S < \gamma_m, \forall S \in \mathcal{B}^{(\ell)}) \to 1$.

Let $\mathcal{F}^{(\ell)} \subset \mathcal{A}_{\text{md}} \cap \mathcal{A}^{(\ell)}$ with $|\mathcal{F}^{(\ell)}| = c_{\text{md}}$. Define

$$\hat{\mathcal{F}}^{(\ell)} = \big\{A \in \mathcal{B}^{(\ell)} \cap \mathcal{F}^{(\ell)} : T_A < \alpha_m\big\}$$

By condition 2.2, $\forall A \in \mathcal{F}^{(\ell)}$,

$$\mathsf{P}(A \in \hat{\mathcal{F}}^{(\ell)}) \geq \mathsf{P}\big(T_A < \alpha_m, T_D \geq \bar{\Phi}(m^{r_1-1}\sqrt{\log m}), \forall D \in \mathcal{D}(A)\big) \geq C_1 \qquad (2.24)$$

Accordingly, define $\hat{\mathcal{X}}^{(\ell)} = \{|\hat{\mathcal{F}}^{(\ell)}| \geq c_{\mathrm{md}}/2\}$, then $\mathsf{P}(\hat{\mathcal{X}}^{(\ell)}) \geq 1 - o(1)$.

On $\hat{\mathcal{X}}^{(\ell)}$, we have

$$\sum_{S \in \mathcal{B}_1^{(\ell)}} I(T_S \leq \hat{t}^{(\ell)}) \geq C c_{\mathrm{md}}$$

Then based on Lemma 2.3, we can conclude that $\mathsf{P}(m_0 \hat{t}^{(\ell)} \geq C^{(\ell)} c_{\mathrm{md}}) \geq 1 - o(1)$ for some constant $C^{(\ell)}$. □

Proof of Theorem 2.2. Let $\mathcal{V}^{(\ell)} = \{S \in \mathcal{B}_0^{(\ell)} : S \subset \mathcal{R}^{(\ell)}\}$ and $\mathcal{W}^{(\ell)} = \{S \in \mathcal{B}_1^{(\ell)} : S \subset \mathcal{R}^{(\ell)}\}$ be the false rejection node set and the rejection node set on layer ℓ, respectively. Define

$$\mathcal{X}_1 = \{S \in \cup_{\ell=2}^{L} \mathcal{W}^{(\ell)} : S \cap \Omega_{\mathrm{st}}^{(1:L)} \neq \emptyset \text{ and } S \cap \Omega_0 \neq \emptyset\}$$

$$\mathcal{X}_2 = \{S \in \cup_{\ell=2}^{L} \mathcal{W}^{(\ell)} : S \cap \Omega_{\mathrm{wk}} \neq \emptyset,\ S \setminus (\Omega_0 \cup \Omega_{\mathrm{wk}}) = \emptyset \text{ and } S \cap \Omega_0 \neq \emptyset\}$$

$$\mathcal{X}_3 = \{S \in \cup_{\ell=2}^{L} \mathcal{W}^{(\ell)} : S \cap \Omega_1 \setminus (\Omega_{\mathrm{wk}} \cup \Omega_{\mathrm{st}}^{(1:L)}) \neq \emptyset \text{ and } S \cap \Omega_0 \neq \emptyset\}$$

Then,

$$\mathsf{P}(\mathcal{X}_1 \neq \emptyset) \leq \mathsf{P}(\mathcal{X}_1 \neq \emptyset | \cap_{\ell=1}^{L} \mathcal{X}^{(\ell)}) \mathsf{P}(\cap_{\ell=1}^{L} \mathcal{X}^{(\ell)}) + \mathsf{P}((\cap_{\ell=1}^{L} \mathcal{X}^{(\ell)})^c) \leq C m^{r_1} o(m^{-r_1}) + o(1) \to 0$$

$$\begin{aligned}
\mathsf{P}(\mathcal{X}_2 \neq \emptyset) &\leq \mathsf{P}(\mathcal{X}_2 \neq \emptyset | \cap_{\ell=1}^{L} \mathcal{X}^{(\ell)}) \mathsf{P}(\cap_{\ell=1}^{L} \mathcal{X}^{(\ell)}) + \mathsf{P}((\cap_{\ell=1}^{L} \mathcal{X}^{(\ell)})^c) \\
&\overset{(a)}{\leq} C m^{r_1} \mathsf{P}\left[X_S \geq \beta_m \middle| S \in \Omega_{\mathrm{wk}} \cup \Omega_0 \right] + o(1) \\
&\leq C m^{r_1} o(m^{-r_1}) + o(1) \to 0
\end{aligned}$$

Here, the inequality (a) is based on Lemma 2.4 (1) and (3). By condition 2.4,

$|\mathcal{X}_3| = o(c_{\text{md}})$, accordingly,

$$\begin{aligned}\mathsf{P}\big(FDP > \alpha + \epsilon\big) \leq & \mathsf{P}(\mathcal{X}_1 \cup \mathcal{X}_2 \neq \emptyset) + \mathsf{P}\left(\frac{\sum_{\ell=1}^{L}\sum_{S\in\mathcal{V}^{(\ell)}}|S|}{\sum_{\ell=1}^{L}\sum_{S\in\mathcal{R}_{\text{node}}^{(\ell)}}|S|} > \alpha+\epsilon, \mathcal{X}_1\cup\mathcal{X}_2 = \emptyset\right) \\ \leq & o(1) + \sum_{\ell=1}^{L}\mathsf{P}\left(\frac{\sum_{S\in\mathcal{V}^{(\ell)}\setminus\mathcal{X}_3}|S|}{\sum_{S\in\mathcal{R}_{\text{node}}^{(\ell)}}|S|} > \alpha+\epsilon+o(1)\right) \to 0\end{aligned}$$

So statement (1) is proved. The statement (2) can be proved in the similar way. □

3

TEAM: A Multiple Testing Algorithm on the Aggregation Tree for Flow Cytometry Analysis

3.1 Introduction

Flow cytometry is a multivariate single-cell assay commonly used to characterize the immune system. A key challenge in flow cytometry is the identification of T cells that are activated by specific antigens such as a particular tumor, bacterial or viral protein. Intracellular cytokine staining (ICS) is often combined with flow cytometry to analyze the antigen-specific T cell immune response. In ICS, cells are first activated with antigen, followed by staining for cell surface molecules that define the cell phenotype, such as CD4 and/or CD8. Cells can be further stained after membrane permeabilization with fluorochrome-labeled monoclonal antibodies specific to protein markers within the cell, such as the IFN-γ, IL-2, and TNF-α cytokines that are only expressed after T cell activation. Hence T cells with high levels of cytokine expressions are likely to be antigen-specific. The flow cytometer quantifies the amount of antibodies bound to the cell and hence the concentration of the protein marker that the antibody is targeting. The expression repertoire of

protein markers is often used to characterize cell types at two levels – the cell surface markers, which define the basic cell type (*e.g.*, CD4+ T cell) and maturational stage (*e.g.*, memory T cells), and the intracellular cytokines, whose expressions define a cell's activation class. Quantification of different activation classes of antigen-specific T cells provides useful biological information such as the likely efficacy of a vaccine (Seder et al., 2008). However, as the relative frequency of antigen-specific cells for any particular antigen is often very low (rarely above 1% and often much lower), several hundred thousand cells per sample are typically evaluated to quantify different subpopulations of antigen-specific T cells. This gives rise to the need for statistical methods that can detect antigen-specific cells but guard against false positives.

Generally speaking, in a typical flow cytometry study, N_1 cells under condition 1 (cohort 1) and N_2 cells under condition 2 (cohort 2) are evaluated; for each cell, p protein marker expressions are collected. Here N_1 and N_2 could range from 10^5 to 10^7, and p could range from fewer than 10 up to 50. The goal is to pinpoint the activated cells. The characterization of cells activated under condition 2 has many applications - for example, in evaluating the immunogenicity of a vaccine. As mentioned above, the activated cells are typically present in very low relative frequencies. Identifying those cells directly using an analytic method is very challenging. Thus, we first identify the sample space regions that are highly enriched for such cells. These regions could still contain some non-activated cells, so further filtering is needed to pinpoint the activated cells. The filtering could be based on other cell protein markers from the domain knowledge or based on sub-analyses of multiple functional markers.

Many existing methods used this strategy to first identify differential density regions. Roederer and Hardy (2001) proposes the frequency gating method (Roederer et al., 2001a) to compare the frequencies between stimulated and reference cells and identify

regions with significant differential frequencies. The reference sample space is divided into bins containing roughly equal numbers of cells. The resulting partition is then applied to the stimulated sample, and the ratio of the stimulated and reference cells in each bin is calculated to derive a normalized chi-square statistic for each bin. Then they use a user-defined threshold to reject those bins with large chi-square statistics. Duong (2013) uses kernel density estimation and chi-square test statistics to identify local distributional differences at any given location in multi-dimensional space. Antoniadis et al. (2015) focuses on identifying one-dimensional differential density regions. They divide the samples into equal length bins and then use the Poisson distribution to model the number of stimulated and reference cells in each bin. Then they used a variance stabilization transformation to normalize the counts and applied existing multiple-testing procedures to identify differential density regions. More recently, Soriano and Ma (2017) proposed a multi-resolution scanning (MRS) method for identifying differential density regions. The method can be viewed as a testing method embedded in the partition tree. MRS sequentially partitions the common support of two distributions into finer and finer bins. On every resolution level, each bin is coupled with a null hypothesis. MRS forms a hypothesis for each large or small bin and asymptotically controls the false discovery rate (FDR) among all these hypotheses.

Pura (2020) model the challenge of pinpointing differential regions as a multiple-testing problem, and propose a new FDR controlling procedure, called TEAM. First, TEAM partitions the multivariate sample space into bins with the finest resolution. It can accommodate different partitioning schemes. Second, TEAM embeds testing on an aggregation tree. On layer 1, within each bin, TEAM tests if the PDF of cohort 2 is higher than that of cohort 1. On higher layers, TEAM will gradually aggregate the accepted bins and test if the aggregated bins harbor differential PDFs. This fine-

resolution to coarse-resolution testing structure not only boosts the testing power but also pinpoints the regions with differential PDFs at the finest possible resolution.

In this chapter, we provide a rigorous theoretical study about the TEAM framework. Specifically, we justify the false discovery control of TEAM through rigorous proof. The rest of the chapter is organized as follows. Section 3.2 introduces how to model the flow cytometry data. Section 3.3 describes the TEAM algorithm. Section 3.4 and 3.5 provides the theoretical justification of the TEAM algorithm. The proofs of Lemmas are shown in Appendix B.

3.2 Model

3.2.1 Sample space partitioning

Let Ω be the multivariate protein marker sample space of the pooled cells (cohort 1 and cohort 2). Consider a partition on Ω, *i.e.*, $\cup_{i=1}^{m}\Omega_i = \Omega$, $\mu(\Omega_i \cap \Omega_j) = 0$ if $i \neq j$. We call $\Omega_1, \ldots, \Omega_m$ the leaf bins (of leaf $1, \ldots, m$). Suppose the leaf bin Ω_i consists of n_i cells. The partition can be constructed in several ways. Two examples include:

- *Adaptive partition.* We order the protein markers from the largest expression variance to the smallest. We choose the marker with the largest sample variance and do a median split of the sample space along this dimension. Within each subsample, we repeat these median splits of the sample space along the dimension with the largest variance. After $\tilde{m}$ times, we will have $m = 2^{\tilde{m}}$ bins. See Roederer et al. (2001b) and Roederer et al. (2001a) for details.
- *Sequential partition.* We order the protein markers from the largest expression variance to the smallest. We first partition the first marker dimension into $\tilde{m}$ bins by its sample quantiles at level $\{1/\tilde{m}\}$, ..., $\{(\tilde{m}-1)/\tilde{m}\}$. Within each bin on the first dimension, we partition the second dimension by its sample quantiles at level $\{1/\tilde{m}\}$, ..., $\{(\tilde{m}-1)/\tilde{m}\}$. We then sequentially partition all

other dimensions until all dimensions are partitioned.

3.2.2 Hypotheses

Let $\tilde{X}_i$ and X_i be the number of cells from cohort 1 and cohort 2 that falls into leaf bin i, respectively. Then $n_i = X_i + \tilde{X}_i$. We know that $\sum_{i=1}^m \tilde{X}_i = N_1$ and $\sum_{i=1}^m X_i = N_2$. Consider the problem with the fixed margins N_1, N_2, and n_i, $i = 1, \ldots, m$. Clearly, $\{X_1, \ldots, X_m\}$ are not mutually independent. However, for any finite K-dimensional vector $(X_{i1}, \ldots, X_{iK})'$, its joint distribution can be well approximated by the product of the mutually-independent binomial distributions with the i-th component $\text{Binom}(n_i, \theta_i)$, where

$$\theta_i = \frac{N_2 \int_{\Omega_i} f_2(y)\,\mathrm{d}y}{N_2 \int_{\Omega_i} f_2(y)\,\mathrm{d}y + N_1 \int_{\Omega_i} f_1(y)\,\mathrm{d}y}. \tag{3.1}$$

Here, f_1 and f_2 are the PDFs of cohort 1 and cohort 2, and N_1 and N_2 are the cell numbers in cohort 1 and cohort 2.

From here, we consider this problem in the probability space after partition. Noteworthy, the new sample space after partition $\tilde{\Omega}$ contains all possible realizations of $X_1, \ldots, X_m$ conditioning on N_1, N_2, and $n_1, \ldots, n_m$. It is different from the original sample space Ω. We also need to define the probability measure conditioning on the partition. Lemma 3.1 characterizes this measure.

Lemma 3.1. *Let $Z_1, \ldots, Z_m$ be a sequence of independent random variables, where Z_i follows* $\text{Binom}(n_i, \theta_i)$. *Suppose $n^{(1)} = \sup_{i=1}^m n_i = n_i + \delta_{n,i}$ with $\sup_{i=1}^m |\delta_{n,i}| = o(n^{(1)})$ and $n^{(1)} = o(N^{1/2})$. For any constant K and vector $(i_1, \ldots, i_K)$, with each element taken without replacement from $\{1, \ldots, m\}$,*

$$\left| \frac{\mathsf{P}(X_{i_1} = x_{i_1}, \ldots X_{i_K} = x_{ik} \mid N_1, N_2, n_{i_1}, \ldots, n_{i_K})}{\prod_{j=1}^K \mathsf{P}(Z_{i_j} = x_{i_j} \mid n_{i_j})} - 1 \right| \leq CK^2 (n^{(1)})^2 / N,$$

for some constant C not depending on K, $n^{(1)}$, m, N_1, N_2, or $(x_{i1}, \ldots, x_{iK})'$.

When we justify the asymptotic properties of TEAM, only joint distributions of finite dimensional $(X_{i_1}, \ldots, X_{i_K})'$ are involved. Therefore, Lemma 3.1 is sufficient.

Let $\theta_0 = N_2/N$. For leaf i, we set up the hypothesis:

$$\mathrm{H}_{\mathrm{nul},i} : \theta_i \leq \theta_0 \quad \text{versus} \quad \mathrm{H}_{\mathrm{alt},i} : \theta_i > \theta_0. \tag{3.2}$$

If $\mathrm{H}_{\mathrm{alt},i}$ is true we call leaf i alternative and null otherwise.

We consider one-sided tests here because they correspond to our analytical goal of locating the activated cells in cohort 2. Clearly, $\int f_s(y)\,\mathrm{d}y = 1$ for both $s \in \{1, 2\}$. By the continuity of f_1 and f_2, if there exists a region where the cohort 2 density is higher, there must exist a region where the cohort 1 density is higher. For our analytical goal, we only need to find regions where the cohort 2 density is higher. Under some rare cases, researchers are interested in identifying regions with differential densities in either direction; then we can first run the one-sided test, and flip the labels of cohort 1 and cohort 2, and run the one-sided test again.

Let $\mathcal{H} = \{1, \ldots, m\}$, $\mathcal{H}_{\mathrm{nul}} = \{i : \mathrm{H}_{\mathrm{nul},i} \text{ is true}\}$, and $\mathcal{H}_{\mathrm{alt}} = \mathcal{H} \setminus \mathcal{H}_{\mathrm{nul}}$. Ideally, we would like to identify where $f_2 > f_1$, *i.e.*, the region $\Omega^+ = \{y \in \Omega : f_2(y) > f_1(y)\}$. With the partitioned bins, the goal is to identify $\hat{\Omega}^+ = \cup_{i \in \mathcal{H}_{\mathrm{alt}}} \Omega_i$. To make sure that $\hat{\Omega}^+$ approximates Ω^+ well, the partition should be fine enough. However, if the partition is too fine and each leaf bin only contains very few cells, the testing power will be low. Discussions on the proper theoretical choice of m and n_i are provided in the Section 3.4.

3.3 Algorithm Description

3.3.1 Step 1: Testing on layer 1

The false discovery proportion (FDP) and false discovery rate (FDR) is defined as

$$\text{FDP} = \frac{\sum_{i\in\mathcal{H}_{\text{nul}}} I(\text{ H}_{\text{nul},i} \text{ is rejected})}{\sum_{i\in\mathcal{H}} I(\text{ H}_{\text{nul},i} \text{ is rejected}) \vee 1}, \quad \text{FDR} = \text{E}(\text{FDP}). \tag{3.3}$$

To control FDR, we propose the following test procedure on the bottom layer. Let $G_{0,i}^{(1)}$ be the complementary cumulative density function (CCDF) of $\text{Binom}(n_i, \theta_0)$. Let $P_{0,i}^{(1)} = G_{0,i}^{(1)}(X_i)$, a random variable close to the P-value. Define the threshold $\hat{c}^{(1)}$ as

$$\hat{c}^{(1)} = \sup\left\{ a_N^{(1)} \le c \le \alpha : c \le \frac{\max\left\{\sum_{i\in\mathcal{H}} I(P_{0,i}^{(1)} < c), 1\right\}}{m} \cdot \alpha \right\}. \tag{3.4}$$

Here $a_N^{(1)} = (m \log m)^{-1}$. If such $\hat{c}^{(1)}$ does not exist, set $\hat{c}^{(1)} = a_N^{(1)}$. We reject $\text{H}_{\text{nul},i}$ if $P_{0,i} \le \hat{c}^{(1)}$. This procedure is very similar to the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995).

3.3.2 Step 2: Aggregation and Testing on higher layers.

Unlike other multiple testing methods, TEAM will continue testing after layer 1. It will aggregate the neighboring accepted leaves to parent nodes and test their parent hypotheses. The underlying assumption of TEAM is that the neighboring leaves of an alternative leaf are also likely to be alternative. This assumption is reasonable under many circumstances, especially when the PDFs f_1 and f_2 are smooth (See Proposition 3.1). Thus, TEAM hierarchically aggregates the neighboring leaves so that their signals can be aggregated and amplified. It is possible that an alternative region spanning multiple leaf bins is missed on low layers but will be identified on higher layers.

Figure 3.1 provides a toy example to illustrate how TEAM works.

- Leaf 1 and leaf 2 are accepted on layer 1, so they are aggregated as a parent node $S_1^{(2)} = \{1, 2\}$. The coupled node hypothesis is

$$\mathrm{H}_{\mathrm{nul},1}^{(2)} : \ \forall\ j \in S_1^{(2)},\ \theta_j \leq \theta_0 \quad \text{versus} \quad \mathrm{H}_{\mathrm{alt},1}^{(2)} : \ \exists\ j \in S_1^{(2)},\ \theta_j > \theta_0.$$

- Leaf hypothesis $\mathrm{H}_{\mathrm{nul},8}^{(1)}$ is rejected on the bottom layer. As a result, leaves 7 and 9 are aggregated on layer 2, so that $S_4^{(2)} = \{7, 9\}$. This aggregation design allows the differential peak to be captured at a lower layer and the differential shoulder to be captured at a higher layer. See the illustrated distributions in Figure 3.1. Aggregating the shoulder areas will likely increase power.
- Leaf 12 is left alone on the bottom layer because no more leaves are left to be aggregated. On any layer, at most 1 node will be left out.

More generally, higher layers of TEAM employ two steps: aggregation and testing.

Aggregation. On layer ℓ, we aggregate the neighboring accepted child nodes on layer $\ell - 1$ into the parent nodes. If the leaf bins are ordinal, the aggregation can be easily performed according to the ordinal rankings. Figure 3.1 illustrates a one-dimensional example. The one-dimensional example can be easily extended to multiple dimensions.

Node hypothesis. After aggregation, new parent nodes $S_i^{(\ell)}$ are formulated. We set up the coupled node hypothesis:

$$\mathrm{H}_{\mathrm{nul},i}^{(\ell)} : \forall\ j \in S_i^{(\ell)},\ \theta_j \leq \theta_0 \quad \text{versus} \quad \mathrm{H}_{\mathrm{alt},i}^{(\ell)} : \exists\ j \in S_i^{(\ell)},\ \theta_j > \theta_0. \tag{3.5}$$

We test these hypotheses on layer ℓ and map the rejections back to the bottom layer with the finest resolution.

Testing. On layer ℓ, suppose there are $m^{(\ell)}$ aggregated nodes on layer ℓ. For $\ell \geq 2$, each node $S_i^{(\ell)}$ is the union of two child nodes on $\ell - 1$, denoted by $S_{i_1}^{(\ell-1)}$ and

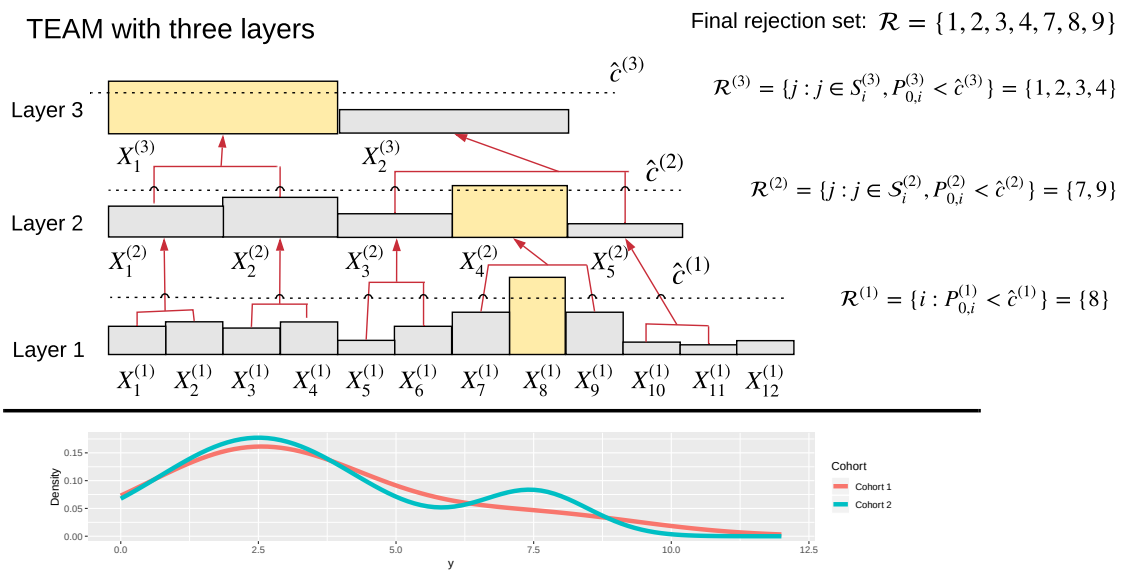


FIGURE 3.1: An illustrating example of TEAM with three layers. The non-rejected bins are aggregated at the beginning of layer 2 and layer 3, and each parent bin is coupled with a parent hypothesis. If a parent hypothesis is rejected, the rejection is mapped back to the bottom layer. For example, at the beginning of layer 2, the non-rejected leaf bin set is $\tilde{\mathcal{H}}^{(2)} = \{1,2,3,4,5,6,7,9,10,11,12\}$, and the parent bin set is $\mathcal{A}^{(2)} = \{\{1,2\},\{3,4\},\{5,6\},\{7,9\},\{10,11\}\}$. On layer 2, the null hypothesis coupled with the parent bin $\{5,6\}$ is rejected. The rejection is mapped to the bottom layer so that $\mathrm{H}_{\mathrm{nul},5}$ and $\mathrm{H}_{\mathrm{nul},6}$ are rejected.

$S_{i_2}^{(\ell-1)}$. Obviously $|S_i^{(\ell)}| = 2^{\ell-1}$. The node bin contains $n_i^{(\ell)} = \sum_{j \in S_i^{(\ell)}} n_j^{(1)}$ samples, out of which $X_i^{(\ell)} = \sum_{j \in S_i^{(\ell)}} X_j^{(1)}$ are from cohort 2. It is easy to see that $X_i^{(\ell)} = X_{i_1}^{(\ell-1)} + X_{i_2}^{(\ell-1)}$, where $X_{i_1}^{(\ell-1)}$ and $X_{i_2}^{(\ell-1)}$ are the number of cohort 2 samples in the bins of its child node $S_{i_1}^{(\ell-1)}$ and $S_{i_2}^{(\ell-1)}$.

Similar to layer 1, we first derive a P-value-like statistic $P_{0,i}^{(\ell)} = G_{0,i}^{(\ell)}(X_i^{(\ell)}; \hat{c}^{(\ell-1)})$. Here $G_{0,i}^{(\ell)}(c; \hat{c}^{(\ell-1)})$ is defined recursively. On layer 1, $G_{0,i}(c; \hat{c}^{(0)})$ is the CCDF of $\mathrm{Binom}(n_i^{(1)}, \theta_0)$. On layer $\ell \geq 2$, define

$$G_{0,i}^{(\ell)}(c; \hat{c}^{(\ell-1)}) = \mathsf{P}_{\theta_0}(Z_1 > c \mid G_{0,i_1}^{(\ell-1)}(Z_1; \hat{c}^{(\ell-2)}) > \hat{c}^{(\ell-1)}, G_{0,i_2}^{(\ell-1)}(Z_2; \hat{c}^{(\ell-2)}) > \hat{c}^{(\ell-1)}), \tag{3.6}$$

where Z_1 and Z_2 independently follows $\text{Binom}(n_{i_1}^{(\ell-1)}, \theta_0)$ and $\text{Binom}(n_{i_2}^{(\ell-1)}, \theta_0)$ respectively, and $Z = Z_1 + Z_2$.

Technically speaking, $P_{0,i}^{(\ell)}$ is not the P-value of $X_i^{(\ell)}$ because the null distribution of $X_i^{(\ell)}$ depends on the entire testing path. However, we can show that $|P_{0,i}^{(\ell)} - \breve{P}_{0,i}|/\breve{P}_{0,i}$ converges to zero in probability, where $\breve{P}_{0,i}$ is the P-value of $X_i^{(\ell)}$ (Lemma 3.8). Thus $P_{0,i}^{(\ell)}$ asymptotically follows $\text{Unif}(0,1)$ under $\text{H}_{\text{nul},i}^{(\ell)}$.

We will reject the node hypothesis $\text{H}_{\text{nul},i}^{(\ell)}$ if $\mathsf{P}_{0,i}^{(\ell)} \le \hat{c}^{(\ell)}(\alpha)$, where

$$\hat{c}^{(\ell)}(\alpha) = \sup\left\{ a_N^{(\ell)} \le c \le \alpha : c \le \frac{\max\left[\sum_{1\le i\le m^{(\ell)}} I\{P_{0,i}^{(\ell)} \le c\}, 1\right]}{m^{(\ell)}} \cdot \alpha \right\}, \tag{3.7}$$

with $a_N^{(\ell)} = (m^{(\ell)} \log m^{(\ell)})^{-1}$. If such $\hat{c}^{(\ell)}$ does not exist, set $\hat{c}^{(\ell)} = a_N^{(\ell)}$.

Here, $a_N^{(\ell)}$ is set at $(m^{(\ell)} \log m^{(\ell)})^{-1}$ because later we will prove

$$\mathsf{P}\left(\sup_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} P_{0,i}^{(\ell)} < a_N^{(\ell)} \right) \le \sum_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} P(P_{0,i}^{(\ell)} < a_N^{(\ell)}) = O(1/\log m) = o(1),$$

where $\mathcal{B}_{\text{nul}}^{(\ell)}$ is the null node set on layer ℓ. In other words, if the threshold is set up at $\hat{c}^{(\ell)}$, the probability of making any false rejections on layer ℓ is negligible. Thus, there is no need to consider smaller cutoffs. Similar bounds are also used in Liu et al. (2013), Liu et al. (2014), Xie and Li (2018).

To map the node-level rejections to the leaves (on layer 1), we adopt an aggressive approach:

$$\text{If } \text{H}_{\text{nul},i}^{(\ell)} \text{ is rejected}, \quad \text{then } \forall\, j \in S_i^{(\ell)}, \text{H}_{\text{nul},j} \text{ are rejected.}$$

For example, in Figure 3.1, we reject $\text{H}_{\text{nul},1}^{(3)}$. Its corresponding node is $S_1^{(3)} = \{1,2,3,4\}$. Then we reject $\text{H}_{\text{nul},1}$, $\text{H}_{\text{nul},2}$, $\text{H}_{\text{nul},3}$, and $\text{H}_{\text{nul},4}$. This aggressive ap-

proach is based on the underlying assumption that the alternative leaves are likely to cluster. See Section 3.3.5 for justification.

The hierarchical structure of TEAM is designed to boost the testing power. On the bottom layer, we test one leaf at a time. When a leaf harbors a strong signal θ_i, it is likely to be rejected on layer 1. On higher layers, leaves are aggregated into larger nodes, and these nodes will be tested collectively. Because the neighboring leaves all have similar signal levels θ_i (based on Condition 3.4 in Section 3.4), some weak-signal leaves have the chance to be aggregated. Therefore, the collective signal of the node will be much stronger, and the node will have a higher chance to be rejected. The rejections will be mapped back to the leaf layer. Thus, compared to just testing the individual leaves on the bottom layer, TEAM will be more powerful.

3.3.3 Stopping rule

To control when TEAM stops, we set up a flag: flag $= 0$ for proceeding and flag $= 1$ for stopping. At the beginning of TEAM, flag $= 0$ and switches to 1 when the stopping rule is satisfied. Here are some examples of stopping rules.

- We set up a predetermined number L as the maximum layer. Then TEAM will stop after L layers.
- After the testing procedure on layer ℓ for any $\ell \geq 2$, we calculate the number of rejections on layer ℓ. If the number is less than a prespecified level, TEAM will stop.
- After the testing procedure on layer ℓ for any $\ell \geq 2$, we calculate the ratio between the rejection number on layer ℓ and on layer $\ell - 1$. If the ratio is below a prespecified level, TEAM will stop.

3.3.4 Pseudocode for TEAM

1. Set flag $= 0$ and $\ell = 1$.
2. On layer 1, define the leaf $S_i^{(1)} = \{i\}$ for $i = 1, \ldots, m^{(1)}$ with $m^{(1)} = m$. Let the rejection set be
$$\mathcal{R}^{(1)} = \{i : \ P_{0,i}^{(1)} \leq \hat{c}^{(1)}\}$$
with $\hat{c}^{(1)}$ defined in (3.4).
3. Check the stopping rule. If it is satisfied, set flag $= 1$ and go to Step 4; otherwise, increase ℓ by 1 and perform the following sub-steps on layer ℓ ($\ell \geq 2$).
   (a) Let $\tilde{\mathcal{H}}^{(\ell)} = \mathcal{H}^{(\ell-1)} \setminus \mathcal{R}^{(\ell-1)}$.
   (b) Based on the predefined aggregation rule, let the node $S_i^{(\ell)} = S_{i_1}^{(\ell-1)} \cup S_{i_2}^{(\ell-1)}$ for $i = 1, \ldots, m^{(\ell)}$. Here $|S_i^{(\ell)}| = 2^\ell - 1$ and $m^{(\ell)} = \lfloor |\mathcal{H}^{(\ell)}|/2^{\ell-1} \rfloor$.
   (c) Set $\mathcal{H}^{(\ell)} = \cup_{i=1}^{m^{(\ell)}} S_i^{(\ell)}$, $\mathcal{H}_{\text{nul}}^{(\ell)} = \mathcal{H}^{(\ell)} \cap \mathcal{H}_{\text{nul}}$, and $\mathcal{H}_{\text{alt}}^{(\ell)} = \mathcal{H}^{(\ell)} \cap \mathcal{H}_{\text{alt}}$.
   (d) Obtain the rejection set
$$\mathcal{R}^{(\ell)} = \{j : j \in S_i^{(\ell)}, \ P_{0,i}^{(\ell)} \leq \hat{c}^{(\ell)}\}.$$
   with $\hat{c}^{(\ell)}$ defined in (3.7).
4. Let the overall rejection set be $\mathcal{R} = \cup_{i=1}^{L} \mathcal{R}^{(\ell)}$. We reject $\text{H}_{\text{nul},i}$ for all $i \in \mathcal{R}$.

3.3.5 Justification of the aggressive rejection rule

On layer ℓ, after we aggregated the leaf bins into larger bins, for their index set $S_i^{(\ell)}$ ($i \in \{1, \ldots, m^{(\ell)}\}$), we define

$$\text{H}_{\text{nul},i}^{(\ell)} : \ \forall \ j \in S_i^{(\ell)}, \theta_j \leq \theta_0 \quad \text{versus} \quad \text{H}_{\text{alt},i}^{(\ell)} : \exists \ j \in S_i^{(\ell)}, \theta_j > \theta_0. \tag{3.8}$$

When $\text{H}_i^{(\ell)}$ is rejected, all $\text{H}_{\text{nul},j}$ with $j \in S_i^{(\ell)}$ are rejected. This rejection rule is very aggressive. Yet, it makes sense under the assumption that null (alternative) hypotheses tend to cluster together along the order. In other words, if $\text{H}_{\text{alt},i}$ is true,

then its neighboring hypotheses are more likely to be alternative. This underlying assumption is reasonable in the setting of identifying differential density regions especially when both f_1 and f_2 satisfy the following regularity conditions.

Lemma 3.2. *Suppose that $f_1(y)$ and $f_2(y)$ are two probability density functions on a closed sample space $\Omega \subset \mathbb{R}^p$ satisfying*

$$0 < M_1 \leq \inf_{s\in\{1,2\},\ y\in\Omega} f_s(y) \leq \sup_{s\in\{1,2\},\ y\in\Omega} f_s(y) \leq M_2, \tag{3.9}$$

and

$$\sup_{s\in\{1,2\},\ y\in\Omega} \|\nabla f_s\|_2 \leq M_3, \quad \textit{where } \nabla f_s = \left(\frac{\partial f}{\partial y_1}, \ldots, \frac{\partial f}{\partial y_p}\right)'. \tag{3.10}$$

Assume the partition $\{\Omega_1, \ldots, \Omega_m\}$ on Ω satisfies

$$\sup_{i=1}^{m} \mu(\Omega_i) \leq M_4/m, \quad \textit{and } \sup_{i=1}^{m} \mathsf{Span}(\Omega_i) \leq M_5/m^{1/p}.$$

For some constant $M_4, M_5 > 0$. Here μ is the Lebesgue measure and $\mathsf{Span}(\Omega_i) = \sup_{y_1,y_2\in\Omega_i}\|y_1 - y_2\|_2$. Then

$$\sup_{i\in\{2,\ldots,m\}} |\theta_i - \theta_{i-1}| \leq O(1/m). \tag{3.11}$$

Lemma 3.2 requires several mild conditions. The first condition requires that f_1 and f_2 are bounded functions with bounded first derivatives. This condition holds for a wide range of functions. The second condition requires the partition to be roughly "even" on the sample space Ω and across all dimensions. The second condition is satisfied with many partition methods, for example, the sequential partition strategy.

Proposition 3.1. *Consider the probability density functions satisfying* (3.9) *and* (3.10)*. With the probability converging to 1, the partition $\{\Omega_1, ..., \Omega_m\}$ generated based on sequential partition strategy satisfies*

$$\sup_{i=1}^{m} \mu(\Omega_i) \leq M_4/m, \quad \textit{and } \sup_{i=1}^{m} \mathsf{Span}(\Omega_i) \leq M_5/m^{1/p}.$$

For some constant $M_4, M_5 > 0$.

3.4 Asymptotic Validity

To simplify, we only consider TEAM to stop at layer L, where L is a constant. First, we introduce the conditions needed to justify the theoretical properties of TEAM. To unify the expression of the conditions on all layers $l \in \{1, \ldots, L\}$, let $\mathcal{B}^{(\ell)} = \{1, \ldots, m^{(\ell)}\}$, the null set $\mathcal{B}^{(\ell)}_{\text{nul}} = \{i : \forall\ j \in S^{(\ell)}_i,\ \theta_j \leq \theta_0\}$, and the alternative set $\mathcal{B}^{(\ell)}_{\text{alt}} = \mathcal{B}^{(\ell)} \setminus \mathcal{B}^{(\ell)}_{\text{nul}}$. Also let $m^{(\ell)}_0 = |\mathcal{B}^{(\ell)}_{\text{nul}}|$, $m^{(\ell)}_1 = |\mathcal{B}^{(\ell)}_{\text{alt}}|$. Then $m^{(\ell)} = m^{(\ell)}_0 + m^{(\ell)}_1$.

Condition 3.1. Assume $m^{(1)}_1 \leq r_2\{m^{(1)}\}^{r_1}$ for some $r_1 < 1/2$, and $r_2 > 0$. Let $n^{(1)} = \sup_{i\in\{1,\ldots,m\}} n_i$. Assume $n_i \geq n^{(1)}[1 - o(\sqrt{\frac{1}{n \log m}})]$ for all $i \in \{1, \ldots, m\}$ and $N^{r_3} \leq n^{(1)} \leq N^{r_4}$ for some constants $\frac{r_1}{p+r_1} < r_3 \leq r_4 < \frac{1-r_1}{1+p-r_1}$, where $N = N_2 + N_1$ is the total number of cells.

Condition 3.2. For all $i \in \{1, \ldots, m\}$, assume $r_5 \leq \theta_i \leq 1 - r_5$ for some constants r_5 satisfying $0 < r_5 < 0.5$.

Condition 3.3. Let $\alpha^{(0)} = +\infty$. For any $1 \leq \ell \leq L$, let

$$\beta = \left\{2(1 - r_1) \log m^{(1)} - \log\log m^{(1)}\right\}^{1/2},\ \gamma = (2 \log m^{(1)})^{1/2},\ \lambda = (\log\log m^{(1)})^{1/2} \tag{3.12}$$

Define

$$\mathcal{G}^{(1)} = \left\{ i : (n^{(1)})^{1/2}(\theta_i - \theta_0) > 2^{-1}\lambda + \{\theta_0(1 - \theta_0)\}^{1/2}\gamma \right\}.$$

For $\ell \geq 2$, define

$$\mathcal{E}_i^{(\ell)} = \{i, i+1, \ldots, i+2^\ell - 1\} \subseteq \{1, \ldots, m\},$$

$$\mathcal{G}^{(\ell)} = \Bigg\{ i \in \{1, \ldots, m^{(1)}\} : \ \forall\ j \in \mathcal{E}_i^{(\ell)}, \\ \lambda + \{\theta_0(1-\theta_0)\}^{1/2}\gamma < (n^{(\ell)})^{1/2}(\theta_j - \theta_0) \leq \sqrt{2\theta_0(1-\theta_0)}\beta - \lambda \Bigg\}. \tag{3.13}$$

Assume for some constant $r_6 > 0$, $s_G^{(\ell)} = |\mathcal{G}^{(\ell)}|$ satisfies

$$s_G^{(\ell)} \geq r_6 \log m^{(1)}.$$

Condition 3.4. The neighboring θ_is are similar such that

$$\sup_{j=2,\ldots,m} |\theta_j - \theta_{j-1}| = o\left(\frac{1}{n^{(1)} \log m^{(1)}}\right).$$

Condition 3.1 assumes that the true alternative bins are sparse. The condition $r_1 < 1/2$ is sufficient to guarantee that in (3.13)

$$2^{-1}\lambda + \{\theta_0(1-\theta_0)\}^{1/2}\gamma < \sqrt{2\theta_0(1-\theta_0)}\beta.$$

Condition 3.1 also specifies a lower bound for the number of the pooled cells in each bin. This number cannot be too small to affect the asymptotic convergence in the individual bin test. Also, because $m^{(1)} = N/n^{(1)}$, it also imposes a upper bound for the number of hypotheses on the bottom layer. The condition $N^{r_3} \leq n^{(1)} \leq N^{r_4}$ is equivalent to the condition $(n^{(1)})^{\frac{1-r_4}{r_4}} \leq m^{(1)} \leq (n^{(1)})^{\frac{1-r_3}{r_3}}$, where $m^{(1)}$ is the number of hypotheses on the bottom layer. Besides, $n_i \geq n^{(1)}[1 - o(\sqrt{\frac{1}{n \log m}})]$ requires the samples in each bin are almost the same. Both the *adaptive partition* and the *sequential partition* described in Section 3.3 satisfy this condition.

Condition 3.2 assumes all θ_i is bounded away from 0 and 1.

Condition 3.3 assumed the existence of clustered signals with certain strength levels. The corresponding signal sets are labelled by $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(L)}$. Those $\mathcal{G}^{(\ell)}$ with smaller ℓ contain signal segments where the individual signal is strong (θ_i is large) but the segment is short; Those $\mathcal{G}^{(\ell)}$ with larger ℓ contain signal segments where the individual signal is weak (θ_i is close to θ_0) but the segment is long. The signal level and segment length in $\mathcal{G}^{(\ell)}$ is designed in a way such that at least a sub-segment with length $2^{\ell-1}$ will stay after $\ell - 1$ layers with a non-negligible probability; and on layer ℓ, this sub-segment will be identified with a probability converging to 1. Compared with the upper bound on the total alternatives $r_2(m^{(1)})^{r_1}$, the lower bound on $s_G^{(\ell)}$ is much smaller. This means that we only require some small numbers of signals to be large.

Conditions 3.1–3.3 are required to prove that the layer-specific FDP of TEAM is consistent. Condition 3.4 is required to prove the overall FDP is consistent. Lemma 2 outlines the condition needed to guarantee

$$\sup_{i \in \{2,\ldots,m\}} |\theta_i - \theta_{i-1}| \leq O(1/m^{(1)}).$$

By Condition 3.1, $m^{(1)} \geq \{n^{(1)}\}^{\frac{1-r_4}{r_4}} > \{n^{(1)}\}^{2/3}$. Therefore, $O(1/m^{(1)}) = o\left(\frac{1}{n^{(1)} \log m^{(1)}}\right)$. In other words, when (3.9),(3.10) and Condition 3.1 hold, Condition 3.4 also holds.

Recall our definitions on the parent null and alternative hypothesis sets

$$\mathcal{B}_{\text{nul}}^{(\ell)} = \{i : \ S_i^{(\ell)} \subseteq \mathcal{H}^{(\ell)}, \forall j \in S_i^{(\ell)}, \ \theta_j \leq \theta_0\}, \quad \mathcal{B}_{\text{alt}}^{(\ell)} = \{i : \ S_i^{(\ell)} \subseteq \mathcal{H}^{(\ell)}, \exists j \in S_i^{(\ell)}, \ \theta_j > \theta_0\}.$$

Based on them, we define the false and true rejection sets

$$\mathcal{V}^{(\ell)} = \{i \in \mathcal{B}_{\text{nul}}^{(\ell)} : \ S_i^{(\ell)} \subseteq \mathcal{R}^{(\ell)}\}, \quad \mathcal{W}^{(\ell)} = \{i \in \mathcal{B}_{\text{alt}}^{(\ell)} : \ S_i^{(\ell)} \subseteq \mathcal{R}^{(\ell)}\}.$$

Also define $\mathcal{U}^{(\ell)} = \mathcal{V}^{(\ell)} \cup \mathcal{W}^{(\ell)}$. Let

$$U^{(\ell)} = |\mathcal{U}^{(\ell)}|, \quad V^{(\ell)} = |\mathcal{V}^{(\ell)}|, \quad W^{(\ell)} = |\mathcal{W}^{(\ell)}|.$$

On layer ℓ, the layer-specific FDP and FDR is defined as

$$\mathrm{FDP}^{(\ell)} = V^{(\ell)}/U^{(\ell)}, \quad , \mathrm{FDR}^{(\ell)} = \mathrm{E}(\mathrm{FDP}^{(\ell)}). \tag{3.14}$$

We will first prove that $\mathrm{FDP}^{(\ell)}$ converges to the desired level α in probability. There are two major technical difficulties. First, for any $i \in \mathcal{B}_{\mathrm{nul}}^{(\ell)}$, we know that $\theta_j \leq \theta_0$. However, we calculate the P-value $P_{0,i}^{(\ell)}$ of $X_i^{(\ell)}$ by assuming $\theta_j = \theta_0$. If lots of θ_js are much smaller than θ_0, many $P_{0,i}^{(\ell)}$ will be too conservative. Fortunately, the following lemma guarantees that this will not happen.

Lemma 3.3. *Define the set*

$$\mathcal{D}_{\mathrm{nul}}^{(1)} = \{i \in \mathcal{H}_{\mathrm{nul}}^{(1)} : \theta_0 - (n^{(1)} \log m^{(1)})^{-1} < \theta_i \leq \theta_0\} \tag{3.15}$$

Then under Condition 3.1,

$$\lim_{N\to\infty} \frac{|\mathcal{D}_{\mathrm{nul}}^{(1)}|}{m_0^{(1)}} = 1.$$

Now we proceed to prove the validity of TEAM. For $\ell = 1$, the proof is similar to those in the existing references on single-layer multiple testing, *e.g.*, Liu et al. (2014), Xie and Li (2018), and Xia et al. (2019). For higher layer $\ell \geq 2$, because the aggregation and testing on layer ℓ depends on the testing results on layer $\ell - 1$, we need to prove the validity by induction.

Theorem 3.1. *Under Conditions 3.1-3.3, on layer ℓ, TEAM satisfies*

$$\lim_{N\to\infty} \mathsf{P}(|FDP^{(\ell)} - \alpha| \leq \epsilon) = 1 \text{ for any } \epsilon > 0, \text{ and } \lim_{N\to\infty} FDR^{(\ell)} = \alpha.$$

Theorem 3.1 shows the layer-specific consistency of $\mathrm{FDP}^{(\ell)}$. However, this does not necessarily lead to the overall consistency of FDP. This is because for any $i \in \mathcal{B}_{\mathrm{alt}}^{(\ell)}$ it is possible that only one or a few $\theta_j > \theta_0$. On the other hand, TEAM

has an aggressive rejection rule – once $\mathrm{H}_{\mathrm{nul}}^{(\ell)}$ is rejected, we will reject all its child leaves $\mathrm{H}_{\mathrm{nul},j}^{(1)}$ for all $j \in S_i^{(\ell)}$. Without additional condition, this aggressive rejection rule will introduce many false positives. However, with Condition 3.4 (the similar neighboring θ_i condition), we show that TEAM will have overall FDP converging to α in probability.

After mapping the rejections on the layer ℓ to the bottom layer, denote by $\mathcal{R}_{\mathrm{nul}}^{(\ell)}$, $\mathcal{R}_{\mathrm{alt}}^{(\ell)}$, and $\mathcal{R}^{(\ell)}$ as follows.

$$\mathcal{R}^{(\ell)} = \{j : i \in \mathcal{U}^{(\ell)}, j \in S_i^{(\ell)}\}, \quad \mathcal{R}_{\mathrm{nul}}^{(\ell)} = \{j : i \in \mathcal{U}^{(\ell)}, j \in S_i^{(\ell)} \cap \mathcal{H}_{\mathrm{nul}}\}, \quad \mathcal{R}_{\mathrm{alt}}^{(\ell)} = \mathcal{R}^{(\ell)} \backslash \mathcal{R}_{\mathrm{nul}}^{(\ell)}.$$

It is easy to see that $|\mathcal{R}^{(\ell)}| = 2^{\ell-1}U^{(\ell)}$. The overall FDP and FDR are defined in terms of the leaves

$$\mathrm{FDP}^{(1:L)} = \frac{\sum_{\ell=1}^{L} |\mathcal{R}_{\mathrm{nul}}^{(\ell)}|}{\sum_{\ell=1}^{L} |\mathcal{R}^{(\ell)}|}, \quad \mathrm{FDR}^{(1:L)} = \mathrm{E}(\mathrm{FDP}^{(1:L)}). \tag{3.16}$$

Theorem 3.2. *Under Conditions 3.1–3.4, the overall false discover rate*

$$\mathit{FDP}^{(1:L)} \text{ converges to } \alpha \text{ in probability.}$$

3.5 Proofs of the Main Results

For simplicity of the proof of the asymptotic properties, let $n_i = n$ for $i = 1, \ldots, m$. Then $n^{(\ell)} = 2^{\ell-1}n$. Recall that $S_i^{(\ell)}$ is the index set coupled with node i on layer ℓ. Further, denote by $\widetilde{G}_i^{(\ell)}(c^{(\ell)}; c^{(\ell-1)})$ the conditional CCDF of $X_i^{(\ell)}$ conditioning on $X_{i_1}^{(\ell-1)} \le c^{(\ell-1)}, X_{i_2}^{(\ell-1)} \le c^{(\ell-1)}$. When $\ell = 1$, $c^{(0)} =: +\infty$. Then $\widetilde{G}_i^{(1)}(c^{(1)}; c^{(0)})$ decays to the marginal complementary CDF of $\mathrm{Binom}(n_i, \theta_i)$.

For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $a_n \le Cb_n$ holds for all sufficiently large n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. If $a_n = O(b_n)$ and $b_n = O(a_n)$, then $a_n \asymp b_n$. If $\lim_{n\to\infty} a_n/b_n = 1$, write $a_n \sim b_n$.

To prove the asymptotic properties of TEAM, we need the following lemmas. Lemma provide a binomial local limit theorem and moderate deviation result for the binomial tail probability. The proofs of Lemmas 3.4 (a) and 3.5 can be found in Chapter 8 of Lesigne (2005). The proofs of the other Lemmas can be found in Section B.1.

Lemma 3.4. *For a* $\text{Binom}(n, \theta)$ *random variable* $\breve{X}$*, whenever* k *satisfies* $|k - n\theta| < c_n n^{2/3}$ *with* $\lim_{n\to} c_n = 0$*, we have*

a)

$$\mathsf{P}(\breve{X} = k) = \frac{1}{\sqrt{2\pi n\theta(1-\theta)}} \exp\left(-\frac{(k-n\theta)^2}{2n\theta(1-\theta)}\right) \cdot (1 + \epsilon_n(k)).$$

with

$$\lim_{n\to\infty} \max_{k:|k-n\theta|<c_n n^{2/3}} |\epsilon_n(k)| = 0$$

b) Consider a random variable $\breve{X}' \sim \text{Binom}(n', \theta)$*, where* n' *satisfies* $|n' - n| \leq n\delta_0(m, n)$ *with* $\delta_0(m, n) = o\left(\sqrt{\frac{1}{n \log m}}\right)$*. Then, if* $|k - n\theta| \leq C\sqrt{n \log m}$ *for some constant* C*, we have*

$$\mathsf{P}(\breve{X}' = k) = \frac{1}{\sqrt{2\pi n\theta(1-\theta)}} \exp\left(-\frac{(k-n\theta)^2}{2n\theta(1-\theta)}\right) \cdot (1 + \epsilon'_n(k)).$$

with

$$\lim_{n\to\infty} \max_{k:|k-n\theta|<C\sqrt{n\log m}} |\epsilon'_n(k)| = 0$$

Lemma 3.5. *Let* $\breve{X}$ *be a* $\text{Binom}(n, \theta)$ *random variable. Suppose that* $\{\tau_n\}$ *is a sequence of real numbers such that* $\lim_{n\to\infty} \tau_n = +\infty$ *and* $\lim_{n\to\infty} \tau_n n^{-1/6} = 0$*. Then*

$$\mathsf{P}(\breve{X} \geq n\theta + \tau_n\sqrt{n\theta(1-\theta)}) \sim \frac{\varphi(\tau_n)}{\tau_n}.$$

Here, $\varphi(\cdot)$ *is the standard normal density function.*

Lemma 3.6. *Consider the* $\mathcal{D}_{\text{nul}}^{(1)}$ *defined in* (3.15). *For* $c^{(\ell-1)} < 1/2$ *and* $b_i^{(\ell)} \geq n_i^{(\ell)}\theta_0$,

$$\lim_{n,m\to\infty} \sup_{S_i^{(\ell)} \subseteq \mathcal{D}_{\text{nul}}^{(1)}} \frac{\left|\widetilde{G}_i^{(\ell)}(b_i^{(\ell)}; c^{(\ell-1)}) - G_{0,i}^{(\ell)}(b_i^{(\ell)}; c^{(\ell-1)})\right|}{G_{0,i}^{(\ell)}(b_i^{(\ell)}; c^{(\ell-1)})} = 0.$$

Lemma 3.7.

$$\sup_{b: b - n_i\theta_0 < \sqrt{2n\log m\theta_0(1-\theta_0)}} \left| \frac{G_{0,i}^{(1)}(b + o(\sqrt{\frac{n}{\log m}}); \hat{c}^{(0)})}{G_{0,i}^{(1)}(b; \hat{c}^{(0)})} - 1 \right| \to 0$$

Lemma 3.8. *Let* $\breve{X}_i \sim \text{Binom}(n_i^{(1)}, \theta_0)$, *with* $i = 1, ..., 2^{\ell-1}$, *and* $\beta_0 = b_0\sqrt{2(1-r_1)\log m}$, *with* $b_0 = \sqrt{\frac{\frac{3}{4} - \frac{r_1}{2}}{1-r_1}} \in (\frac{1}{\sqrt{2(1-r_1)}}, 1)$. *Then,*

a) Let $b_k = n^{(k)}\theta_0 + \sqrt{n^{(k)}\theta_0(1-\theta_0)}\beta_0$, *we have,*

$$\max_{k=1,...,\ell-1} \sup_{b^{(\ell)} \in [n^{(\ell)}\theta_0, n^{(\ell)}\theta_0 + \sqrt{n^{(\ell)}\theta_0(1-\theta_0)}\gamma]} \frac{\mathsf{P}(\sum_{i=1}^{2^{\ell-1}} \breve{X}_i > b^{(\ell)}, \sum_{i=1}^{2^{k-1}} \breve{X}_i > b_k)}{\mathsf{P}(\sum_{i=1}^{2^{\ell-1}} \breve{X}_i > b^{(\ell)})} \to 0 \tag{3.17}$$

b) Let $\hat{b}_i^{(\ell)}$ *be the value s.t.* $G_{0,i}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)}) = \hat{c}^{(\ell)}$, *and*

$$\hat{\tau}_i^{(\ell)} = \frac{\hat{b}_i^{(\ell)} - n_i^{(\ell)}\theta_0}{\{n_i^{(\ell)}\theta_0(1-\theta_0)\}^{1/2}}. \tag{3.18}$$

For $k = 1, ..., \ell - 1$, *when* $P(\hat{\tau}_i^{(k)} \geq \beta_0) \to 1$, *we have*

$$|P_{0,i}^{(\ell)} - \breve{P}_{0,i}|/\breve{P}_{0,i} \to 0, \text{ and} \tag{3.19}$$

$$|G_{0,i}^{(\ell)}(b_i^{(\ell)}; \hat{c}^{(\ell-1)}) - G_{0,i}^{(\ell)}(b_i^{(\ell)}; \hat{c}^{(0)})|/G_{0,i}^{(\ell)}(b_i^{(\ell)}; \hat{c}^{(0)}) \to 0 \tag{3.20}$$

Lemma 3.9. *For every layer* $\ell = 1, ..., L$, *define the set* $\mathcal{X}^{(\ell)}$ *based on a sequence* $\delta_2(m) = o(1)$ *as*

$$\mathcal{X}^{(\ell)} = \left\{ x : \left| \frac{\sum_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}) - \sum_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} P(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)})}{\{\sum_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} P(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)})\}} \right| > \delta_2(m) \right\} \tag{3.21}$$

If the FDR control holds on layer $1, ..., \ell-1$*, and for* $h = 1, ..., \ell$*, there exists constant* $C^{(h)}$*, s.t.*

$$\mathsf{P}(m_0 \hat{c}^{(h)} \geq C^{(h)} \log m) \to 1, \tag{3.22}$$

then under Conditions 3.1 and 3.2, the following three statements hold.

a) Exists constant $\delta_2(m) = o(1)$*, s.t.* $\mathsf{P}(\cap_{h=1}^{\ell} \mathcal{X}^{(h)}) \geq 1 - o(1)$.

b) On $\cap_{h=1}^{\ell} \mathcal{X}^{(h)}$*, there exists some* C *such that*

$$\hat{c}^{(\ell)} \leq C(m^{(\ell)})^{r_1 - 1}. \tag{3.23}$$

c) Recall β *defined in* (3.12) *and* $\hat{\tau}_i^{(\ell)}$ *defined in* (3.18)*. On* $\cap_{h=1}^{\ell-1} \mathcal{X}^{(h)}$*,*

$$\hat{\tau}_i^{(\ell-1)} \geq \beta(1 + o(1)). \tag{3.24}$$

Lemma 3.10. *Under Conditions 3.1 and 3.2, on* $\mathcal{X}$ *defined in* (3.21)*, for all* $\ell \in \{1, \ldots, L\}$*,*

$$\frac{m_0^{(\ell)} G_0^{(\ell)}(\hat{b}^{(\ell)}; \hat{b}^{(\ell-1)})}{\max\left\{\sum_{i=1}^{m^{(\ell)}} I(X_i^{(\ell)} > \hat{b}^{(\ell)}), 1\right\}} = \alpha\{1 + o(1)\}. \tag{3.25}$$

Proof of Proposition 3.1. It is suffice to prove that

$$\sup_{i=1}^{m} P(\mathsf{Span}_j(\Omega_i) \notin [\frac{1}{2M_2 m^{1/p}}, \frac{2}{M_1 m^{1/p}}]) \to o(m^{-1}) \ \forall j \in \{1, ..., p\} \tag{3.26}$$

Here, $\mathsf{Span}_j(\Omega_i)$ is the length of the sample partition Ω_i on dimension j.

On the jth dimension, we conduct the sample quantile partition on all of the $m^{\frac{j-1}{p}}$ sets generated from the previous dimensions. Each of the set contains $N_{(j)} = \frac{N}{m^{(j-1)/p}}$ many cells.

We first consider the $N_{(j)}$ samples $Y_1, ..., Y_{N_{(j)}}$ drawn independently from $UNIF(0, 1/M_1)$, with m and $n_{(j)} = N_j/m^{1/p}$ satisfied the condition 3.1. After we prove those samples

drawn from the uniform distribution satisfy (3.26), we would extend the proof to the case when samples are drawn from the distribution with density $f : [0, 1] \to [M_1, M_2]$

Define a sequence of random variables $0 \leq Q_1 \leq ... \leq Q_{m^{1/p}-1} = 1$ as the $1/m^{1/p}, ..., (m^{1/p} - 1)/m^{1/p}$th sample quantile locations. We also define $Q_0 = 0$ and $Q_{m^{1/p}} = 1$.

If exists $C > 0$, s.t. $P(Q_1 > \frac{2n_{(j)}}{M_1 N_{(j)}}) > \frac{C}{m}$, let $S_1 = \sum_{k=1}^{N_{(j)}} I(Y_k \leq \frac{2n_{(j)}}{M_1 N_{(j)}})$, then

$$P(S_1 \leq n_{(j)}) \geq P(Q_1 > \frac{2n_{(j)}}{M_1 N_{(j)}}) > \frac{C}{m}$$

However, based on Chernoff bound, we have

$$P(S_1 \leq n_{(j)}) \leq P\big(|S_1 - 2n_{(j)}| > n_{(j)}/3\big) = o(m^{-1})$$

Which is a contradiction. Thus, we have

$$P(Q_1 - Q_0 > \frac{2n_{(j)}}{M_1 N_{(j)}}) = o(m^{-1})$$

and similarly,

$$P(Q_{m^{1/p}} - Q_{m^{1/p}-1} > \frac{2n_{(j)}}{M_1 N_{(j)}}) = o(m^{-1})$$

Since for $i = 2, ..., m^{1/p} - 1$,

$$
\begin{aligned}
& P(Q_i - Q_{i-1} \geq \frac{2n_{(j)}}{M_1 N_{(j)}}) \\
=& \int_0^{\frac{1}{M_1} - \frac{2n_{(j)}}{M_1 N_{(j)}}} P(Q_i - Q_{i-1} \geq \frac{2n_{(j)}}{M_1 N_{(j)}} | Q_{i-1} = y) P(Q_{i-1} = y) dy \\
=& \int_0^{1 - \frac{2n_{(j)}}{N_{(j)}}} N_{(j)} \binom{N_{(j)} - 1}{(i-1)n_{(j)} - 1} y^{(i-1)n_{(j)} - 1} (1-y)^{N_{(j)} - (i-1)n_{(j)}} \\
& \sum_{s=0}^{n_{(j)}} \binom{N_{(j)} - (i-1)n_{(j)}}{s} \left(\frac{\frac{2n_{(j)}}{N_{(j)}}}{1-y} \right)^s \left(\frac{1 - y - \frac{2n_{(j)}}{N_{(j)}}}{1-y} \right)^{N_{(j)} - (i-1)n_{(j)} - s} dy \\
=& \sum_{s=0}^{n_{(j)}} \frac{N_{(j)}!}{s!(N_{(j)} - s)!} \left(\frac{2n_{(j)}}{N_{(j)}} \right)^s \left(1 - \frac{2n_{(j)}}{N_{(j)}} \right)^{N_{(j)} - s} \\
=& P(S_1 \leq n_{(j)}) = o(m^{-1})
\end{aligned}
$$

Thus,

$$
m \sup_{i=1}^{m^{1/p}} P((Q_i - Q_{i-1}) \geq \frac{2n_{(j)}}{M_1 N_{(j)}}) = o(1)
$$

Let $f(y) = \{N_2 f_2(y) + N_1 f_1(y)\}/(N_2 + N_1)$. It is easy to see that $M_1 \leq f(y) \leq M_2$.

When $N_{(j)}$ samples $Y'_1, ..., Y'_{N_{(j)}}$ are independently drawn from the distribution with density $f : [0, 1] \to [M_1, M_2]$, we have

$$
P(Q'_i - Q'_{i-1} \geq \frac{2n_{(j)}}{M_1 N_{(j)}}) = \int_0^{1 - \frac{2n_{(j)}}{M_1 N_{(j)}}} P(Q'_i - Q'_{i-1} \geq \frac{2n_{(j)}}{M_1 N_{(j)}} | Q'_{i-1} = y') P(Q'_{i-1} = y') dy'
$$

For any $y'_0 \in [0, 1]$, there always exists corresponding $y_0 = F(y'_0)/M_1 \in [0, \frac{1}{M_1}]$, s.t. $M_1 P(Q_{i-1} = y_0) = P(Q'_{i-1} = y'_0)/f(y'_0)$.

When $y'_0 > F^{-1}\big[1 - \frac{2n_{(j)}}{M_1 N_{(j)}}(m^{1/p} - i + 1)\big]$, we have

$$
n_{(j)} \leq (m^{1/p} - i + 1) n_{(j)} \beta_{i,U}(y'_0) \leq (m^{1/p} - i + 1) n_{(j)} \beta_{i,F}(y'_0)
$$

Here, $\beta_{i,U}(y_0') = \frac{2n_{(j)}/[N_{(j)}M_1]}{1-F(y_0')}$, and $\beta_{i,F}(y_0') = \frac{F(y_0'+2n_{(j)}/[N_{(j)}M_1])-F(y_0')}{1-F(y_0')}$.

Let,

$$W_{i,F}(y_0') = \sum_{k=1}^{(m^{1/p}-i+1)n_{(j)}} I(Y_k' \in [y_0', y_0' + \frac{2n_{(j)}}{M_1 N_{(j)}}]) \sim \text{Binom}\big[(m^{1/p}-i+1)n_{(j)}, \beta_{i,F}(y_0')\big]$$

$$W_{i,U}(y_0') = \sum_{k=1}^{(m^{1/p}-i+1)n_{(j)}} I(Y_k \in [y_0, y_0 + \frac{2n_{(j)}}{M_1 N_{(j)}}]) \sim \text{Binom}\big[(m^{1/p}-i+1)n_{(j)}, \beta_{i,U}(y_0)\big]$$

Then

$$\begin{aligned} P(Q_i' - Q_{i-1}' \geq \frac{2n_{(j)}}{M_1 N_{(j)}} | Q_{i-1}' = y_0') &= P(W_{i,F}(y_0') \leq n_{(j)}) \\ &\leq P(W_{i,U}(y_0') \leq n_{(j)}) \\ &= P(Q_i - Q_{i-1} \geq \frac{2n_{(j)}}{M_1 N_{(j)}} | Q_{i-1} = y_0) \end{aligned}$$

When $y_0' \leq F^{-1}\big[1 - \frac{2n_{(j)}}{M_1 N_{(j)}}(m^{1/p}-i+1)\big]$, the corresponding $y_0 = F(y_0')/M_1 \leq \frac{1-\frac{2n_{(j)}}{M_1 N_{(j)}}(m^{1/p}-i+1)}{M_1}$, and

$$\begin{aligned} &\mathsf{P}(Q_{i-1}' = y_0') \\ \leq &C\mathsf{P}\left[\sum_{i=1}^{N_{(j)}} I\left(Y_k \geq \frac{1-\frac{2n_{(j)}}{M_1 N_{(j)}}(m^{1/p}-i+1)}{M_1}\right) \leq (m^{1/p}-i+1)n_{(j)}\right] \\ = &o(m^{-1}) \text{ (Chernoff bound)} \end{aligned}$$

Thus,

$$
\begin{aligned}
&\mathsf{P}(Q'_i - Q'_{i-1} \geq \frac{2n_{(j)}}{M_1 N_{(j)}}) \\
=&\int_0^{F^{-1}\left[1-\frac{2n_{(j)}}{M_1 N_{(j)}}(m^{1/p}-i+1)\right]} \mathsf{P}(Q'_i - Q'_{i-1} \geq \frac{2n_{(j)}}{M_1 N_{(j)}} | Q'_{i-1} = y')\mathsf{P}(Q'_{i-1} = y')dy' \\
&+ \int_{F^{-1}\left[1-\frac{2n_{(j)}}{M_1 N_{(j)}}(m^{1/p}-i+1)\right]}^{1-\frac{2n_{(j)}}{M_1 N_{(j)}}} \mathsf{P}(Q'_i - Q'_{i-1} \geq \frac{2n_{(j)}}{M_1 N_{(j)}} | Q'_{i-1} = y')\mathsf{P}(Q'_{i-1} = y')dy' \\
\leq& o(m^{-1}) + \mathsf{P}(Q_i - Q_{i-1} \geq \frac{2n_{(j)}}{M_1 N_{(j)}}) \\
=& o(m^{-1})
\end{aligned}
$$

Then,

$$\sup_i \mathsf{P}((Q'_i - Q'_{i-1}) \geq \frac{2n_{(j)}}{M_1 N_{(j)}}) = o(m^{-1})$$

Similarly, we can show that

$$\sup_i \mathsf{P}((Q'_i - Q'_{i-1}) \leq \frac{n_{(j)}}{2M_2 N_{(j)}}) = o(m^{-1})$$

□

Proof of Theorem 3.1. For $\breve{X}_i^{(\ell)} \sim \text{Binom}(n_i^{(\ell)}, \theta_i)$, we define $\widetilde{G}_i^{(\ell)}(b; \hat{c}^{(\ell-1)})$ recursively:

$$\widetilde{G}_i^{(\ell)}(b; \hat{c}^{(\ell-1)}) = \mathsf{P}(\breve{X}_i^{(\ell)} > b \mid G_{0,i}^{(\ell)}(\breve{X}_{i_1}^{(\ell-1)}; \hat{c}^{(\ell-2)}) \geq c^{(\ell-1)}, G_{0,i}^{(\ell)}(\breve{X}_{i_2}^{(\ell-1)}; \hat{c}^{(\ell-2)}) \geq c^{(\ell-1)})$$

With $\tilde{G}_i^{(1)}(b; \hat{c}^{(0)}) = \mathsf{P}(\breve{X}_i^{(1)} > b)$.

Let $\hat{b}_i^{(\ell)}$ be the value s.t. $G_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)}) = \hat{c}^{(\ell)}$. Then the random variable $\text{FDP}^{(\ell)}$

can be decomposed to the product of four parts.

$$\mathrm{FDP}^{(\ell)} = \frac{\sum_{i\in\mathcal{B}_{\mathrm{nul}}^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)})}{\sum_{i\in\mathcal{B}_{\mathrm{nul}}^{(\ell)}} \widetilde{G}_i^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)})} \times \frac{\sum_{i\in\mathcal{B}_{\mathrm{nul}}^{(\ell)}} \widetilde{G}_i^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)})/m_0^{(\ell)}}{\hat{c}^{(\ell)}}$$

$$\times \frac{\hat{c}^{(\ell)}}{\max(\sum_{1\le i\le m^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}), 1)/m^{(\ell)}} \times \frac{m_0^{(\ell)}}{m^{(\ell)}} \tag{3.27}$$

Based on (3.27), in order to prove

$$\lim_{N\to+\infty} \mathsf{P}(|\mathrm{FDP}^{(\ell)} - \alpha| \le \epsilon) = 1, \tag{3.28}$$

we only need to prove

$$\mathsf{P}\left\{\left|\frac{m_0^{(\ell)}}{m^{(\ell)}} - 1\right| > \epsilon\right\} \to 0, \quad \text{as } N\to\infty. \tag{3.29}$$

$$\mathsf{P}\left\{\left|\frac{\sum_{i\in\mathcal{B}_{\mathrm{nul}}^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)})}{\sum_{i\in\mathcal{B}_{\mathrm{nul}}^{(\ell)}} \widetilde{G}_i^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)})} - 1\right| > \epsilon\right\} \to 0, \quad \text{as } N\to\infty. \tag{3.30}$$

$$\mathsf{P}\left\{\left|\frac{\sum_{i\in\mathcal{B}_{\mathrm{nul}}^{(\ell)}} \widetilde{G}_i^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)})}{m_0^{(\ell)}\hat{c}^{(\ell)}} - 1\right| > \epsilon\right\} \to 0, \quad \text{as } N\to\infty. \tag{3.31}$$

$$\mathsf{P}\left\{\left|\frac{m^{(\ell)}\hat{c}^{(\ell)}}{\max(\sum_{1\le i\le m^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}), 1)} - \alpha\right| > \epsilon\right\} \to 0, \quad \text{as } N\to\infty. \tag{3.32}$$

We prove (3.29) – (3.32) by induction.

a) First, let $\ell = 1$.

a1) (3.29) holds by Condition 3.1.

a2) Now we prove (3.31).

Given

$$\sum_{i\in\mathcal{B}_{\text{nul}}^{(1)}} \widetilde{G}_i^{(1)}(\hat{b}_i^{(1)};\hat{c}^{(0)}) - m_0^{(1)}\hat{c}^{(\ell)} =$$

$$\sum_{i\in\mathcal{D}_{\text{nul}}^{(1)}} \left\{\widetilde{G}_i^{(1)}(\hat{b}_i^{(1)};\hat{c}^{(0)}) - \hat{c}^{(\ell)}\right\} + \sum_{i\in\mathcal{E}_{\text{nul}}^{(1)}} \left\{\widetilde{G}_i^{(1)}(\hat{b}_i^{(1)};\hat{c}^{(0)}) - \hat{c}^{(\ell)}\right\}, \quad (3.33)$$

where $\mathcal{E}_{\text{nul}}^{(1)} = \mathcal{B}_{\text{nul}}^{(1)} \setminus \mathcal{D}_{\text{nul}}^{(1)}$. By Lemma 3.6 and Lemma 3.3, we have

$$\sum_{i\in\mathcal{D}_{\text{nul}}^{(1)}} \left\{\widetilde{G}_i^{(1)}(\hat{b}_i^{(1)};\hat{c}^{(0)})\right\} = m_0^{(1)}\hat{c}^{(\ell)}(1+o(1)). \quad (3.34)$$

By Lemma 3.3,

$$0 \le \sum_{i\in\mathcal{E}_{\text{nul}}^{(1)}} \left\{\hat{c}^{(\ell)} - \widetilde{G}_i^{(1)}(\hat{b}_i^{(1)};\hat{c}^{(0)})\right\} \le |\mathcal{E}_{\text{nul}}^{(1)}|\hat{c}^{(\ell)} = o\{m^{(1)}\hat{c}^{(\ell)}\}. \quad (3.35)$$

Combining (3.33), (3.34), and (3.35), we will have (3.31).

a3) Based on Lemma 3.9 and (3.34), to prove (3.30), it suffices to show that for some $C^{(1)} > 0$,

$$\mathsf{P}\left\{m_0^{(1)}\hat{c}^{(\ell)} \ge C^{(1)}\log m^{(1)}\right\} \to 1. \quad (3.36)$$

Recall $\mathcal{G}^{(1)}$ defined in (3.13) and $s_G^{(1)} = |\mathcal{G}^{(1)}|$. Take a subset $\mathcal{F}^{(1)} \subseteq \mathcal{G}^{(1)}$, such that $s_F^{(1)} = |\mathcal{F}^{(1)}| = r_6 \log m^{(1)}$. For any $i \in \mathcal{F}^{(1)}$, because $\theta_i(1-\theta_i) \le 1/4$,

$$n^{(1)}\theta_i - \{n^{(1)}\theta_i(1-\theta_i)\}^{1/2}\lambda \ge n^{(1)}\theta_0 + \{n^{(1)}\theta_0(1-\theta_0)\}^{1/2}\gamma.$$

By (3.7), $b_{\max} := n^{(1)}\theta_0 + \{n^{(1)}\theta_0(1-\theta_0)\}^{1/2}\gamma \ge \hat{b}_i^{(1)}\{1+o(1)\}$ for all $i = 1, ..., m$. Combined with Lemma 3.5, $\widetilde{G}_i^{(1)}(b_{\max};\hat{c}^{(0)}) > 1 - \{\log m^{(1)}\}^{-1}(\log\log m^{(1)})^{-1/2}$. Then, by

Lemma 3.1,

$$\sum_{i\in\mathcal{F}^{(1)}} P(P_{0,i}^{(1)} \le \alpha_N^{(1)}) \ge r_6 \log m^{(1)}(1+o(1)). \tag{3.37}$$

By Markov's inequality,

$$P\big(\big|\sum_{i\in\mathcal{F}^{(1)}} I(P_{0,i}^{(1)} < \alpha_N^{(1)}) - \sum_{i\in\mathcal{F}^{(1)}} P(P_{0,i}^{(1)} < \alpha_N^{(1)})\big| \ge \sqrt{\log m \log\log m}\big) = o(1)$$

Thus, with probability converge to 1,

$$\sum_{i\in\mathcal{F}^{(1)}} I(P_{0,i}^{(1)} < \hat{c}^{(1)}) \ge \sum_{i\in\mathcal{F}^{(1)}} I(P_{0,i}^{(1)} < \alpha_N^{(1)}) \ge r_6 \log m^{(1)}(1+o(1))$$

Combined with Lemma 3.10, there exists constant $C^{(1)}$, s.t.

$$\mathsf{P}\left[m_0^{(1)}\hat{c}^{(\ell)} \ge \left\{C^{(1)}\log m^{(1)}\right\}\right] \ge 1 - o(1).$$

a4) Combining Lemma 3.10 and (3.29), (3.32) holds.

b) Now we assume (3.29) – (3.32) hold for layer $1, 2, \ldots, \ell-1$, and immediately (3.28) holds for layer $1, 2, \ldots, \ell-1$. We will prove (3.29) – (3.32) hold for layer ℓ.

b1) First, we prove (3.29).

Denote by $V^{(k)}$ and $R^{(k)}$ the numbers of false discoveries and total rejections on layer k. Then the total number of true discoveries is $R^{(k)} - V^{(k)}$.

Take any $\tilde{\alpha} > \alpha$, we have

$$\mathsf{P}(\cap_{k=1}^{\ell-1}\{\boldsymbol{x} : \mathrm{FDP}^{(k)} < \tilde{\alpha}\}) \to 1. \tag{3.38}$$

When $\cap_{k=1}^{\ell-1}\mathrm{FDP}^{(k)} \le \tilde{\alpha}$ holds, $V^{(k)} \le \tilde{\alpha}R^{(k)}$. Given $m_1^{(1)} = o(m^{(1)})$, it follows that

$$m_0^{(\ell)}/m^{(\ell)} \ge 1 - o(1).$$

Combined with (3.38), we can get (3.29).

b2) For (3.31), let

$$\mathcal{D}_{\text{nul}}^{(\ell)} = \left\{ i \in \{1, \ldots, m_0^{(\ell)}\} : \forall\ j \in S_i^{(\ell)}, \theta_0 - (n \log m^{(1)})^{-1} < \theta_j < \theta_0 \right\}.$$

By the proof of (3.29) in Part b1), we know that

$$|\mathcal{D}_{\text{nul}}^{(\ell)}| \geq \frac{|\mathcal{D}_{\text{nul}}^{(1)}| - 2^{\ell-1} m_1^{(1)}}{2^{\ell-1}}.$$

Combined with $|\mathcal{D}_{\text{nul}}^{(1)}| \sim m^{(1)}$, $m_0^{(\ell)} \sim m^{(\ell)} \sim \frac{m^{(1)}}{2^{\ell-1}}$, we have $|\mathcal{D}_{\text{nul}}^{(\ell)}| \sim m_0^{(\ell)} \sim m^{(\ell)}$. Now following the similar argument in a2) we can prove (3.31) on layer ℓ.

b3) Now we prove (3.30).

Following the similar argument as in a3), we know that to prove (3.30), it suffices to prove that for some constant $C^{(\ell)} > 0$,

$$\mathsf{P}\left\{ m_0^{(\ell)} \hat{c}^{(\ell)} \geq C^{(\ell)} \log m^{(1)} \right\} \to 1. \tag{3.39}$$

For any $i' \in \mathcal{G}^{(\ell)}$ and $1 \leq k \leq \ell$, define,

$$\mathcal{J}_{i'}^{(k)} = \{ i : S_i^{(k)} \subseteq \mathcal{E}_{i'}^{(\ell)} \}.$$

Then based on condition 3.3, exists a constant $C > 0$, s.t.

$$\begin{aligned}
&\mathsf{P}\big\{ |\mathcal{J}_{i'}^{(\ell)}| \geq 1 \text{ and } \exists i \in \mathcal{J}_{i'}^{(\ell)} s.t. X_i^{(\ell)} > \hat{b}_i^{(\ell)} \big\} \\
\geq &\prod_{j \in \mathcal{E}_{i'}^{(\ell)}} \mathsf{P}\big\{ n^{(1)}\theta_0 + \{n^{(\ell)}\theta_0(1-\theta_0)\}^{1/2}\gamma/2^{\ell-1} \leq X_j^{(1)} \leq n^{(1)}\theta_0 + \{n^{(\ell-1)}\theta_0(1-\theta_0)\}^{1/2}\beta/2^{\ell-2} \big\} \\
\geq &C,
\end{aligned}$$

uniformly for any $i' \in \mathcal{G}^{(\ell)}$.

Let $\mathcal{F}^{(\ell)}$ be the subset of $\mathcal{G}^{(\ell)}$ such that $s_F^{(\ell)} = |\mathcal{F}^{(\ell)}| = r_6 \log m^{(1)}/2^L$, and elements in $\mathcal{F}^{(\ell)}$ are mutually disjoint. Also let

$$\widetilde{\mathcal{F}}^{(\ell)} = \left\{ i' : i' \in \mathcal{F}^{(\ell)} \text{ and there exists some } i \text{ s.t. } i \in \mathcal{J}_{i'}^{(\ell)} \text{ and } X_i^{(\ell)} > \hat{b}_i^{(\ell)} \right\}.$$

By Chebyshev's inequality

$$\mathsf{P}\left\{ \left| |\widetilde{\mathcal{F}}^{(\ell)}| - \sum_{i \in \mathcal{F}^{(\ell)}} \mathsf{P}(i \in \widetilde{\mathcal{F}}^{(\ell)}) \right| > \left\{ \sum_{i \in \mathcal{F}^{(\ell)}} \mathsf{P}(i \in \widetilde{\mathcal{F}}^{(\ell)}) \right\}^{1/2} (\log m^{(1)})^{1/4} \right\}$$

$$\leq \frac{\sum_{i \in \mathcal{F}^{(\ell)}} \mathsf{P}(i \in \widetilde{\mathcal{F}}^{(\ell)}) \{1 - \mathsf{P}(i \in \widetilde{\mathcal{F}}^{(\ell)})\}}{(\log m^{(1)})^{1/2} \sum_{i \in \mathcal{F}^{(\ell)}} \mathsf{P}(i \in \widetilde{\mathcal{F}}^{(\ell)})} \leq (\log m^{(1)})^{-1/2}.$$

Define

$$\mathcal{X}_2 = \cup_{h=1}^{\ell-1} \mathcal{X}^{(h)} \cap \{ |\widetilde{\mathcal{F}}^{(\ell)}| \geq c(\log m^{(1)}) \}. \tag{3.40}$$

Then $\mathsf{P}(\mathcal{X}_2) \geq 1 - o(1)$.

On $\mathcal{X}_2$, we have,

$$\sum_{i \in \tilde{\mathcal{F}}^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}) \geq \sum_{i \in \tilde{\mathcal{F}}^{(\ell)}} I(P_{0,i}^{(\ell)} < \alpha_N^{(\ell)}) \geq C \log m^{(1)}$$

Combined with Lemma 3.10, we can get (3.39).

b4) Finally, by combining Lemma 3.10 and (3.29), we have (3.32).

□

Proof of Theorem 3.2. $\forall i$, let $j_i^{(\ell)} = \arg\sup_{j \in S_i} \theta_j$. If $i \in \mathcal{W}^{(\ell)}$, then $\theta_{j_i^{(\ell)}} > \theta_0$. Let

$$\mathcal{X}_3 = \left\{ \boldsymbol{x} : \inf_{\ell \in \{1,\ldots,L\}} \inf_{i \in \mathcal{W}^{(\ell)}} \theta_{j_i^{(\ell)}} \geq \theta_0 + C_i^{(\ell)} (n^{(1)} \log m)^{-1} \right\} \tag{3.41}$$

If for some $C_i^{(\ell)}$ (may depending on i and ℓ), $\mathsf{P}(\mathcal{X}_3) \to 1$, then by Condition 3.4, we know that

$$\mathsf{P}(\forall \ell = 1, \ldots, L \text{ and } i \in \mathcal{W}^{(\ell)}, \text{we have } S_i^{(\ell)} \subseteq \mathcal{H}_{\text{alt}}) \to 1.$$

Subsequently,

$$\mathsf{P}\big\{\forall \ell = 1, \ldots, L,\ |\mathcal{R}_{\text{nul}}^{(\ell)}| = 2^{\ell-1} V^{(\ell)}\big\} \to 1.$$

Combined with Theorem 3.1, we can prove the overall $\text{FDP}^{1:L}$ is constant to α.

Now it suffices to prove that for some $C_i^{(\ell)}$, $\Pr(\mathcal{X}_3) \to 1$.

For any i with $S_i^{(\ell)} \cap \mathcal{H}_{\text{alt}} \neq \emptyset$, consider the event $\theta_{j_i^{(\ell)}} - \theta_0 \leq (n^{(1)} \log m)^{-1} \varepsilon_N$, with some $\varepsilon_N \to 0$. $\forall j \in S_i^{(\ell)}$, $\theta_j \leq \theta_{j_i^{(\ell)}}$. Then by the similar arguments as Lemma 3.9 and Lemma 3.5,

$$\begin{aligned}
&\mathsf{P}(X_i^{(\ell)} > \hat{b}_i^{(\ell)} \mid S_i^{(\ell)} \cap \mathcal{H}_{\text{alt}} \neq \emptyset, \theta_{j_i^{(\ell)}} - \theta_0 \leq (n^{(1)} \log m)^{-1} \varepsilon_N, \mathcal{X}) \\
\leq\ &\mathsf{P}(X_i^{(\ell)} > \hat{b}_i^{(\ell)} \mid \forall j \in S_i^{(\ell)}, \theta_j = \theta_{j_i^{(1)}} \leq (n^{(1)} \log m)^{-1} \varepsilon_N, \mathcal{X}) \\
\leq\ &C m^{r_1 - 1}
\end{aligned}$$

Given there would be at most Cm^{r_1} many alternative hypothesis on each layer and $r_1 < 1/2$,

$$\mathsf{P}(i \in \mathcal{W}^{(\ell)}, \theta_{j_i^{(\ell)}} - \theta_0 \leq (n^{(1)} \log m)^{-1} \varepsilon_N, \mathcal{X}) \leq o(1).$$

And accordingly,

$$\mathsf{P}(\mathcal{X}_3) \geq 1 - o(1).$$

□

4

Localizing Rare-Variant Association Regions via Multiple Testing Embedded in an Aggregation Tree

4.1 Introduction

Rare variant (RV) disease association analysis has been an important topic in genetics and genomics. Because single variant association analysis suffers a lack of power (Bansal et al., 2010; Lee et al., 2014), a common approach to increase power is to test the collective effect of all RVs within prefixed regions by burden tests (Li and Leal, 2008; Asimit et al., 2012; Morris and Zeggini, 2010), variant component tests (Neale et al., 2011; Wu et al., 2011), or their combinations(Lee et al., 2012). The prefixed regions could be genes (Petrovski et al., 2017; Povysil et al., 2019), functional domains(Gelfman et al., 2019), or sliding fixed-length windows (Bhatia et al., 2010; Katsumata and Fardo, 2020). However, those prefixed-region-based methods are sub-optimal when the pre-fixed regions are too long or short. For example, when causal RVs are localized in a shorter region, testing the collective effect of the RVs in this shorter region is more powerful than testing over a longer pre-fixed region. Another

example is that, when the pre-fixed region (such as a fixed-size sliding window) is too short, the analysis may exclude adjacent regions containing disease-associated RVs, leading to a false negative for this region.

Inspired by this finding, the varying-window methods (Li et al., 2019; Kanoungi et al., 2020) scan subregions with different lengths to optimize the power of detecting disease-associated RVs. Compared with the fixed-region methods, the varying-window methods are generally more powerful. However, in practice, the varying-window methods often output long regions containing many neutral RVs; thus, they are less useful in pinpointing the functionally important RVs in the downstream analysis. Moreover, searching regions with different lengths is computationally intensive; thus, these methods usually take long.

We propose a new method that powerfully pinpoints the short regions containing the disease-associated RVs. The method dynamically aggregates and tests genomic regions with varying lengths for their disease association, called DATED (Dynamic Aggregation and Tree-Embedded testing). DATED adopts a finer-to-coarse hierarchical testing strategy: given the previous layer's testing results, DATED dynamically defines the testing regions and performs multiple testing on the current layer, and then passes the results to the next layer. Compared with the varying-window methods, DATED has much better precision: it outputs much shorter regions that carry fewer neutral RVs in both numerical and empirical studies (Section 4.3).

Although DATED has a hierarchical testing framework, it is fundamentally different from the existing hierarchical methods. The existing methods include the methods developed for the ordered hypotheses (Dmitrienko and Tamhane, 2013), or the hypotheses coupled with the prefixed nodes on the trees (Yekutieli, 2008; Soriano and Ma, 2017; Li et al., 2020) or directed acyclic graphs (DAGs) (Meijer and Goeman, 2015; Guo et al., 2018). Their main differences are the following. First, DATED

is customized for RV-disease-association studies. It dynamically defines the smallest testing regions based on the sample size so that studies with large sample size have higher pinpointing resolution and specificity. In contrast, most existing hierarchical testing methods only have improved sensitivity when sample sizes increase but not precision. Second, DATED's aggregation and testing system is dynamic: the latter layer's DGS, hypotheses, and testing rules depend on the previous layers' results. In contrast, the existing hierarchical testing methods test static nodes and hypotheses. The dynamic strategy will improve testing speed and precision. Third, DATED adopts a new type I error measure, node-FDR, to satisfy the need to test RV and disease associations. Node-FDR is region-level false discovery rate (FDR) weighted by the proportions of neutral RVs: For two genetic sets both containing the disease-associated RVs, the one with lower proportions of neutral RVs will have lower node-FDR and higher disease susceptibility. By adopting node-FDR as a type I error measure, DATED prioritizes finding those regions with higher concentrations of disease-associated RVs and increases the pinpointing accuracy.

We demonstrated that DATED outperforms other competing methods under various numerical settings: DATED usually has on-par performance on the domain-level sensitivity compared with the state-of-the-art methods; however, its RV-level precision is much higher. We also apply DATED to a study of amyotrophic lateral sclerosis (ALS) to pinpoint the regions containing ALS-associated RVs. DATED successfully identified regions in the well-known ALS genes (such as *SOD1* and *TARDBP*) and a new region in *EPGS* (MIM: 615068) on chromosome 18. This region in *EPGS* is missed by the competing methods. In addition, the regions detected by DATED tend to be shorter; these regions also harbor RVs with higher exposure rates in ALS patients.

4.2 Materials and Methods

4.2.1 Notations

DATED adopts a dynamic and hierarchical testing structure. It splits the whole exome or genome into small regions called leaves, tests their disease associations, and hierarchically aggregate these regions and test again. Here, we introduce the notations required for describing the algorithm.

Leaves and nodes

Consider a case-control study with N_0 controls, N_1 cases, and in total J observed qualified RVs in D functional domains. We will partition these D functional domains into $m^{(1)}$ leaves. In this study, we only focus on the functional domain regions. However, we can perform a similar partition on the whole genome region, including the non-coding regions.

To optimize the testing power, we define a leaf as a region where M subjects carry the RV mutations. Here, M is called leaf size. It is a much smaller number compared to N_0 and N_1. When $M = 1$, each leaf contains only one qualified RV. In practice, we recommend $M \geq 5$, and a detailed M selection criterion is discussed later. With genome domain annotations, we can further optimize the definition of leaves by forcing the breakdown of leaves at the beginning and the end of the domains. Thus, a leaf will not spread two domains. If fewer than M subjects carry the RV mutations in the ending leaf of a domain, the ending leaf will be combined with the preceding leaf in the same domain. See Algorithm 6 in Appendix C.

Suppose the partition ends up with $m^{(1)}$ leaves, denoted by $\{1\}, \ldots, \{m^{(1)}\}$. The set of all leaves are denoted by $\mathcal{B}^{(1)} = \{\{i\} : i \in [m^{(1)}]\}$, where $[m^{(1)}]$ means the set $\{1, \ldots, m^{(1)}\}$.

DATED has multiple layers. On higher layers (layer $\ell \geq 2$), it will aggregate the accepting nodes from the previous layer to form new nodes. A node S is a union of one or more leaves. A node represents a genome region with qualified RVs. Thus, leaves are a type of node. For any node S, the set of its qualified RVs is denoted by $\mathcal{V}(S)$, $\mathcal{V}(S) \subset [J]$.

On layer ℓ, if a node is the union of two or more nodes on layer $\ell - 1$, it is called a qualified node; the set of all qualified nodes is denoted by $\mathcal{B}^{(\ell)}$.

Node hypotheses and P-values

For any leaf $\{i\}$, we call it null if

for all $j \in \mathcal{V}(\{i\})$, the qualified RV j is not associated with the disease;

otherwise, we call it alternative. Denote the set of null leaves by $\mathcal{B}_0^{(1)}$ and the set of alternative leaves by $\mathcal{B}_1^{(1)}$.

Similarly, for any node S, we define the node hypothesis H_S:

$$H_{S,0} : \forall\ i \in S, \{i\} \in \mathcal{B}_0^{(1)} \quad \text{versus} \quad H_{S,1} : \exists\ i \in S, \{i\} \in \mathcal{B}_1^{(1)}. \tag{4.1}$$

For simplicity, for leaf hypothesis $H_{\{i\}}$, we denote them as H_i. H_i is null if $\{i\} \in \mathcal{B}_0^1$ and alternative if $\{i\} \in \mathcal{B}_1^{(1)}$.

To test these hypotheses, we need to summarize their P-values. DATED is flexible to accommodate the existing RV test statistics such as the burden tests (Asimit et al., 2012; Morgenthaler and Thilly, 2007; Li and Leal, 2008), the variance-component tests (Wu et al., 2011; Pan, 2009), or the omnibus tests (Lee et al., 2012). For fast computation, we use the following two efficient approaches.

(1) **Lancaster's mid-P correction for the Fisher's exact test (FL).** First, we use the Fisher's exact test to test the association between the leaf and the

disease. Then we use the Lancaster's mid-P correction to calculate the P-value. The Lancaster's mid-P correction can improve the power of the Fisher's Exact test while still controlling the type I error (Lancaster, 1961; Biddle and Morris, 2011).

(2) **Efficient score statistics with saddle point approximation (SS).** This approach is to calculate the collapsing score statistic based on a logistic regression model and then use the saddle point approximation to adjust for the statistic's tail distribution. (Daniels, 1954; Dey et al., 2017) The method is computationally efficient and robust when the number of subjects carrying the minor alleles is small.

More details are provided in Appendix C. After leaf P-values are defined, the node P-values are aggregations of the leaf P-values. For any node S, we calculate its P-value by

$$T_S = \bar{\Phi}(\sum_{i \in S} \bar{\Phi}^{-1}(T_i)/\sqrt{|S|}), \tag{4.2}$$

where $\bar{\Phi}(\cdot)$ is the complementary cumulative density function (CCDF) of the standard Gaussian distribution.

Error criterion of DATED: node-FDR

When a node is the union of multiple leaves, we might not be able to dichonomize its null and alternative status. For example, suppose node $S = \{1, 2\}$ with leaf $\{1\}$ is null and leaf $\{2\}$ is alternative. Then S is $1/2$ alternative. More generally, for any node S, let $S_1 = S \cap \{i : H_i \text{ is alternative}\}$. Then $\theta_S = {}^{|S_1|}/_{|S|}$. We call S is θ_S-alternative or $(1 - \theta_S)$-null.

To address this fractional null/alternative status of nodes, We introduce a new type

I error measure called node-FDR:

$$\text{node-FDP} = \frac{\sum_{S\in\mathcal{R}}\{1-\theta_S\}}{|\mathcal{R}|\vee 1}, \quad \text{node-FDR} = \text{E}(\text{node-FDP}), \tag{4.3}$$

where $\mathcal{R}$ is the set of the rejected nodes. Clearly, rejecting a node S will contribute $1-\theta_S$ to the numerator and 1 to the denominator. To control node-FDR, DATED will prioritize identifying the nodes containing higher proportions of alternative leaves. When the alternative leaves are highly concentrated in a small region, DATED will reject this small region rather than a more extended region containing many additional null leaves.

4.2.2 DATED: Dynamic and Hierarchical Testing

Algorithm overview

The purpose of DATED is to pinpoint the genomic regions with disease-associated qualified RVs. It not only aims to identify the disease-associated regions but also to prioritize the region with more associated RVs and thus higher disease susceptibility. Towards this aim, DATED uses a fine-to-course testing strategy. It first tests the leaves, the regions with the finest resolution. Then, it hierarchically aggregates leaves into larger node regions, and tests their disease associations. The general algorithm is summarized in Algorithm 1. The details of remarked steps (setting leaf P-value cutoffs, aggregation, and setting node P-value cutoffs) are provided below. Note that the input of DATED depends on the tuning parameter selection. We describe it at the end of the steps. To illustrate the implementation of Algorithm 1, we provide a toy example in Figure 4.1.

Algorithm 1: Dynamic hierarchical testing

Data: Maximum Layer L, leaf P-values $(T_1, \ldots, T_{m^{(1)}})$, and the desired FDR level α

Result: The set of rejected nodes $\mathcal{R}$.

Derive the leaf P-value cutoff $\hat{t}^{(1)}$; *// Setting leaf P-value cutoff*

Set the rejection set: $\mathcal{R} = \{\{i\} : T_i < \hat{t}^{(1)}, i \in [m^{(1)}]\}$;

for $\ell \in \{2, \ldots, L\}$ **do**

  Derive the aggregated node set $\mathcal{B}^{(\ell)}$; *// Aggregation*

  for $S \in \mathcal{B}^{(\ell)}$ **do** Derive node P-values as in (4.2);

  Derive the node P-value cutoff $\hat{t}^{(\ell)}$; *// Setting node P-value cutoffs*

  Update the rejection set: $\mathcal{R} = \mathcal{R} \cup \{S \in \mathcal{B}^{(\ell)} : T_S < \hat{t}^{(\ell)}\}$;

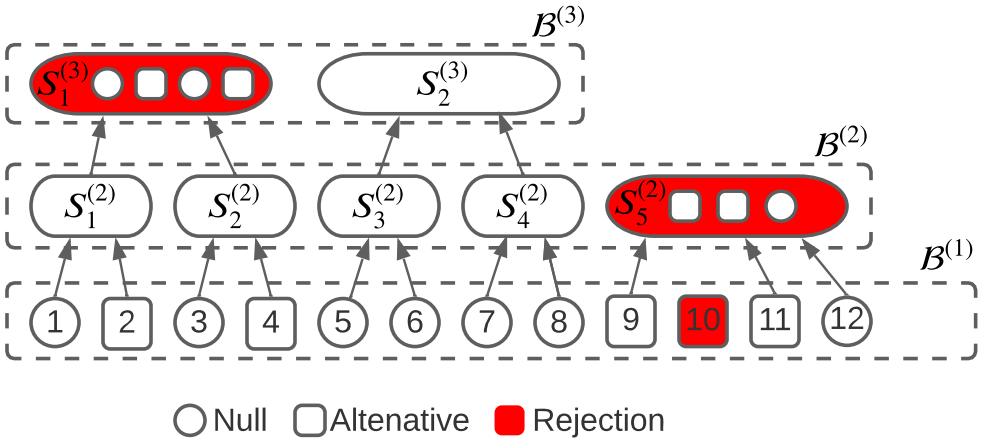


FIGURE 4.1: A toy example of a three-layer DATED analysis. Node i on layer ℓ is denoted by $S_i^{(\ell)}$. On layer 1, we reject leaf $\{10\}$. On layer 2, we aggregate the neighboring accepted leaves into 5 nodes, from which we reject $S_5^{(2)} = \{9, 11, 12\}$. On layer 3, we reject $S_1^{(3)}$. Clearly, $S_{10}^{(1)}$ is 1-alternative, $S_5^{(2)}$ is $2/3$-alternative, and $S_1^{(3)}$ is $1/2$-alternative. The empirical node-FDR of this example is $5/18$.

Setting leaf P-value cutoff

On layer 1 (the leaf layer), let P-value cutoff be

$$\hat{t}^{(1)} = \sup\left\{\alpha^{(1)} \le t \le \alpha : \frac{m^{(1)}t}{\{\sum_{i\in[m^{(1)}]} I(T_i \le t)\} \vee 1} \le \alpha\right\}, \tag{4.4}$$

where $\alpha^{(1)} = 1/\{m^{(1)}(\log m^{(1)})^{1/2}\}$. Here, $a \vee b$ means taking the maximum of a and b. It is easy to see that as long as $\alpha \ge 1/(\log m^{(1)})^{1/2}$, $\hat{t}^{(1)}$ must exist because $\alpha^{(1)}m^{(1)} = 1/(\log m^{(1)})^{1/2} \le \alpha$. In practice, if $\hat{t}^{(1)}$ does not exist, we set $\hat{t}^{(1)} = \alpha^{(1)}$. We reject H_i if the leaf P-value $P_i < \hat{t}^{(1)}$. This testing procedure is almost equivalent to the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995). Existing studies have shown that it asymptotically controls FDR under α (Liu et al., 2013; Xie and Li, 2018). Notably, on layer 1, all leaves are either 1-alternative or 0-alternative. Thus, node-FDR will be equivalent to the traditional FDR. In other words, the testing procedure also asymptotically controls node-FDR.

Aggregation

Aggregation happens on any later layer ℓ with $\ell \ge 2$. The purpose of aggregation is to construct larger candidate regions to test for their associations. Those regions' subregions have already been tested on previous layers and were not significant.

On layer ℓ with $\ell \ge 2$, let $\mathcal{B}' = \mathcal{B}^{(\ell-1)} \setminus \mathcal{R}$. This is a temporary set that contains all the accepted nodes on layer $\ell - 1$. Along with their order, we aggregate the neighboring nodes in $\mathcal{B}'$ within the functional domain into a new node. Since a functional domain is more likely to perform a uniform function, we do not aggregate two nodes in different functional domains. If the last node in a functional domain is left alone, we aggregate it with the proceeding new node. For details, see Algorithm 7 in Appendix C.

On layer ℓ, we only aggregate the accepting nodes on layer $\ell - 1$; thus, the nodes are dynamic. Correspondingly, the node hypotheses are also dynamic. Therefore, testing those dynamic hypotheses is challenging. We need to adaptively set the node P-value cutoff to make the tests valid and powerful.

Setting node P-value cutoffs

We already discussed the leaf P-value cutoff $\hat{t}^{(1)}$. On higher layers $\ell \geq 2$, the node P-value cutoffs depend on the cutoffs on the previous layers. This is because we are testing dynamic hypotheses, and thus the cutoffs are also dynamic.

On layer $\ell \geq 2$, there may exist nodes that contain both null and alternative leaves. We call them mixed nodes. Here, we let d_1 and d_2 denote the number of alternative and null leaves in mixed nodes. We define $\kappa_\ell(t) = E[\sum_{S \in \mathcal{R}_m^{(\ell)}}(1 - \theta_S)]$ as the expected false discoveries of the mixed nodes on this layer, where $\mathcal{R}_m^{(\ell)}$ is the set of rejected mixed nodes under node P-value cutoff t. For a mixed node composed of d_1 alternative and d_2 null leaves, its expected false discoveries is denoted by κ_{ℓ,d_1,d_2}. The $\kappa_\ell(t)$ can be viewed as a weighted sum of the κ_{ℓ,d_1,d_2}: $\kappa_\ell(t) = \sum_{d_1,d_2} \omega_{d_1,d_2}\kappa_{\ell,d_1,d_2}$. Thus, we can get an estimator $\hat{\kappa}_\ell(t)$ by using the sampling method to estimate the κ_{ℓ,d_1,d_2} and the weight ω_{d_1,d_2}. See Algorithm 2 for details.

On layer ℓ, we define the node P-value cutoff

$$\hat{t}^{(\ell)}(\alpha) = \sup\left\{\alpha^{(\ell)} \leq t \leq \alpha : \frac{\sum_{h=1}^{l-1}\{m^{(h)}\hat{t}^{(h)} + \hat{\kappa}_h(\hat{t}^{(h)})\} + m^{(\ell)}t + \hat{\kappa}_\ell(t)}{\left\{\sum_{h=1}^{\ell-1}|\mathcal{R}^{(h)}| + \sum_{S \in \mathcal{B}^{(\ell)}} I(T_S \leq t)\right\} \vee 1} \leq \alpha\right\}, \tag{4.5}$$

where $\alpha^{(\ell)} = (m^{(\ell)} \log m^{(\ell)})^{-1}$, $m^{(\ell)} = |\mathcal{B}^{(\ell)}|$. If such $\hat{t}^{(\ell)}$ does not exist, set $\hat{t}^{(\ell)} = \alpha^{(\ell)}$.

Remark 4.1. Given the desired FDR α, define $\tilde{P}_1$ as a set of leaf P-values link to the rejected hypotheses on layer 1: $\tilde{P}_1 = \{T_k : T_k \leq \hat{t}^{(1)}\}$. Based the definition of the

Algorithm 2: Algorithm to estimate the expected false rejection number among those nodes with at least one alternative leaf on layer ℓ.

Data: t: rejection threshold;
B: sampling times;
$\hat{P}_0$(resp. $\hat{P}_1$): Estimated set of P-values from null leaves(resp. alternative leaves).// *Remark 4.1*
Result: $\hat{\kappa}_\ell(t)$
$\hat{\mathcal{V}} = \{i : T_k \leq T_{(\sqrt{m})}\}$; $\hat{\mathcal{S}}^{(\ell)} = \{S \in \mathcal{B}^{(\ell)} : S \cap \hat{\mathcal{V}} \neq \emptyset, S \setminus \hat{\mathcal{V}} \neq \emptyset\}$;// *Remark 4.2*
Set $\hat{\mathcal{S}}^{(\ell)}_{d_1,d_2} = \{S \in \hat{\mathcal{S}}^{(\ell)} : |S \cap \hat{\mathcal{V}}| = d_1, |S \setminus \hat{\mathcal{V}}| = d_2\}$, the weight can be estimated by $\hat{\omega}_{d_1,d_2} = |\mathcal{S}^{(\ell)}_{d_1,d_2}|$;
Set $\mathbb{D} = \{d_1, d_2 \in \mathbb{Z}^+ : d_1 + d_2 \leq 3 \times 2^{l-2}\}$;
for $b \in \{1, \ldots, B\}$ **do** // *Remark 4.3*
  for $(d_1, d_2) \in \mathbb{D}$ **do**
    $A_1 = \{d_1$ elements sampled from set $\hat{P}_1$ with replacement.$\}$;// *Remark 4.4*
    $A_2 = \{d_2$ elements sampled from set $\hat{P}_0$ with replacement.$\}$;
    $\tilde{T}_{b,d_1,d_2} = \sum_{T \in A_1 \cup A_2} \bar{\Phi}^{-1}(T)/\sqrt{d_1 + d_2}$;
$\hat{\kappa}_{\ell,d_1,d_2} = \frac{d_2}{B(d_1+d_2)} \sum_{b=1}^{B} \mathbb{I}(\tilde{T}_{b,d_1,d_2} < t)$;
$\hat{\kappa}_\ell(t) = \sum_{(d_1,d_2)\in\mathbb{D}} \hat{\omega}_{d_1,d_2} \hat{\kappa}_{\ell,d_1,d_2}$;

FDR, $\tilde{P}_1$ may contain $\alpha|\tilde{P}_1|$ P-values from null. Thus, the estimated set of P-values from alternative leaves $\hat{P}_1$ is defined as the smallest $\lfloor(1-\alpha)|\tilde{P}_1|\rfloor$ P-values in set $\tilde{P}_1$. If $\lfloor(1-\alpha)|\tilde{P}_1|\rfloor = 0$, we would set $\hat{P}_1 = T_{(1)}$. The estimated set of P-values from null leaves $\hat{P}_0$ is a set of P-values not in $\hat{P}_1$.

Remark 4.2. $\mathcal{S}^{(\ell)}_m$ is a set of empirical mixed nodes and $\hat{V}$ is a set of potentially disease-associated leaves. A node S is empirically mixed if it contains both potential disease-associated leaves and potential neutral leaves. Here, we say a leaf is potentially disease-associated if its corresponding P-value is the smallest $\sqrt{m}$ P-values. DATED assumes the number of disease-associated leaves is $o(\sqrt{m})$. Thus, the number of potentially disease-associated leaves is asymptotically larger than the exact number, and the number of empirical mixed nodes tends to be larger than the exact number of the mixed nodes.

Remark 4.3. The B is the preset times of sampling for the empirical estimation of the correction factor $\kappa_{(\ell)}(t)$. Larger B can provide better estimation and hence better node-FDR control. Based on our numerical experiment, DATED can achieve node-FDR control when we set $B = 100$.

Remark 4.4. The P-values of potentially disease-associated leaves are sampled from the set of disease-significant P-values ($\hat{P}_1$). This sampling approach leads to a smaller estimated P-value in each empirical mixed node. Thus, the $\hat{\kappa}_\ell(t)$ estimated from this approach is larger than the true GS level false rejections.

Tuning parameter selection

The maximum layer L and leaf size M are viewed as tuning parameters.

- We first set M_0 as the smallest value so that we can identify at least one leaf on layer 1. To ensure the stability of the statistical testing, we suggest $M_0 \geq 5$. Denote the number of RVs in domain d by V_d. Because most qualified nodes are formed by aggregating two child nodes on the previous layer, very likely, a null domain d will remain till the layer $\lceil \log_2(V_d/M_0) \rceil$. Since most of the domains are null domains, to ensure 50 nodes on layer L, we set $L = \lceil \log_2(V'/M_0) \rceil$, where V' is the length of domain that ranks the 50 largest domain among all $\{V_d : d \in [D]\}$.
- We set the leaf size M as the largest value to keep the same L.

4.2.3 Simulation Studies

We evaluate the performance of DATED under a wide range of simulation settings.

To mimic the structure of RVs nested in domains, we randomly select 2000 domains from the $34,553$ domains on the human genome in the ALS European ancestry subcohort. Within these 2000 domains, a total of $J = 16,543$ missense RVs are

observed.

Out of the 2000 domains, we randomly select 20 domains with more than 30 RVs to be dense pathogenic domains (DPDs) and 20 additional domains to be sparse pathogenic domains (SPDs). For a DPD, the pathogenic region starts from the beginning of the domain and covers the region with $\min(20, L_d/2)$ observed RVs, where L_d is the number of observed RVs on the DPD. Within this pathogenic region, π of them are randomly selected as pathogenic RVs. In the simulation, we vary $\pi \in \{20\%, 40\%, 60\%, 80\%\}$ to evaluate the performance of DATED under different settings. For an SPD, we randomly select one RV in it to be pathogenic. In total, there are 40 pathogenic domains harboring 97, 171, 243 and 317 pathogenic RVs when $\pi = 20\%$, 40%, 60% and 80%, respectively.

For an individual k, denote its RV vector by $\boldsymbol{G}_k = (G_{k,1}, \ldots, G_{k,J})^T$ and $G_{k,j} \sim \text{Bern}(\rho_j)$ is an indicator about whether subject k carries the mutation of RV j. When ρ_j is small, it is close to the minor allele frequency. We simulate $\log_{10} \rho_j$ from $\text{Unif}(-3, -1.7)$ so that ρ_j ranges from 0.001 to 0.02. The binary phenotype of the individual k is independently generated based on the model

$$\text{logit}\mathsf{P}(Y_k = 1 | Z_{k,1}, Z_{k,2}, \boldsymbol{G}_k) = \gamma_0 + \eta Z_{k,1} + \eta Z_{k,2} + \boldsymbol{G}_k^T \boldsymbol{\beta}.$$

Here, $Z_{k,1}$ and $Z_{k,2}$ are two non-genetic covariates with $Z_{k,1} \sim N(0,1)$ and $Z_{k,2} \sim \text{Bern}(0.5)$, respectively. Their coefficients are η, varying in different settings with $\eta \in \{0, 0.5\}$. The other set of coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T$ represents the RV effect size; we set β_j as a decreasing function of ρ_j, $\beta_j = -c \log_{10} \rho_j$. We set $c = 2.5$ for pathogenic RVs in the DPDs and $c = 7.5$ for those in the SPDs. We set the intercept γ_0 so that the disease prevalence is 1%.

Based on the 8 different combinations of $\pi \in \{20\%, 40\%, 60\%, 80\%\}$ and $\eta \in \{0, 0.5\}$ (covariate), we generated 8 large population, each with 2 million individuals. We conduct the numerical studies with 100 repetitions under each population. In each

repetition, we randomly draw N_1 cases and N_0 controls. We let $N_1 = 1000$ and vary $N_0 \in \{1000, 2000, 3000\}$ to check the performance of DATED when the number of controls increases.

When applying DATED to the simulated datasets, we used both the FL and SS statistics (See Section 4.2.1). Under all scenarios, the M_0 is set as 5 since DATED can always make rejections on this value. According to the tuning parameter selection criterion, the number of layers L is set as 3, and the leaf size M is 9. We also check the robustness of DATED when the leaf sizes vary in $\{5, 7, 9, 11\}$. The number of layers L is set according to the tuning parameter selection criterion: for $M \in \{5, 7, 9\}$, $L = 3$; for $M = 11$, $L = 2$.

We compared the performance of the DATED with competing methods, including the domain-based collapsing (DC) method (Gelfman et al., 2019) with the FL and SS statistics, and the SGANG (Li et al., 2019) with their suggested burden (B), SKAT (S), and SKAT-O (O) statistics and default tunning parameters. The desired domain-FDR for the DC method, FWER for SCANG, and node-FDR for the DATED are all set at 5%. The results yielded from these methods are evaluated using the average node-FDP, RV-FDP, RV-sensitiviy, domain-FDP, and domain-sensitivity. The node-FDP is defined in (4.3), and the rest of measures are defined below:

$$\text{RV-FDP} = \frac{\text{Number of the identified neutral RVs}}{\text{Total number of the identified RVs}};$$

$$\text{RV-Sensitivity} = \frac{\text{Number of the identified pathogenic RVs}}{\text{Total number of the pathogenic RVs}};$$

$$\text{Domain-FDP} = \frac{\text{Number of the identified neutral domains}}{\text{Total number of the identified domains}};$$

$$\text{Domain-Sensitivity} = \frac{\text{Number of the identified pathogenic domains}}{\text{Total number of the pathogenic domains}}.$$

Here, an RV is called "identified" if it is located in one of the identified regions; a domain is called "identified" if any of its RVs, regions, or itself is called significant

by the algorithms. We also adopt RV-level and domain-level F1 scores as an overall measurement of the method performance. High F1 score indicates low FDP and high sensitivity. The RV-level and domain-level F1 scores are defined below:

$$\text{RV-F1} = \frac{2(1-(\text{RV-FDP})) \times \text{RV-Sensitivity}}{1-(\text{RV-FDP})+\text{RV-Sensitivity}};$$

$$\text{Domain-F1} = \frac{2(1-\text{Domain-FDP}) \times \text{Domain-Sensitivity}}{1-\text{Domain-FDP}+\text{Domain-Sensitivity}}.$$

4.2.4 Application to an amyotrophic lateral sclerosis (ALS) study

We apply DATED to ALS data. The goal of this study is to pinpoint the pathogenic RV regions. To achieve this goal, an ALS data with either whole-exome sequencing (WES) or whole-genome sequencing (WGS) samples were generated and preprocessed by Columbia University Precision Medicine (Gelfman et al., 2019).

The pre-processed ALS dataset contains 10138 European-ancestry samples, where 2634 of them are cases, and 7504 of them are controls. We observed 289503 qualifying RVs spanning 34553 functional domains. The average minor allele frequencies (MAFs) are 0.16% and 0.12% among the cases and the controls, respectively.

We apply DATED on this dataset with the node-FDR level 5%. We used both FL and SS statistics, where the SS statistics are generated from model (C.1) with no non-genetic covariates. Based on the tuning parameter selection criterion, we set $M = 7$ and $L = 5$.

We also apply SCANG and DC to the ALS dataset to compare their results with DATED. For SCANG, we used the default tunning parameter setting and considered three types of statistics: burden (SCANG-B), SKAT (SCANG-S), and SKAT-O (SCANG-O). Following Li et al. (Li et al., 2019), the minimum and maximum numbers of variants in the sliding windows in SCANG are set as 1 and 161, respectively. For DC, we consider both FL (DC-FL) and SS (DC-SS) statistics. The desired

FEWR for SCANG and the domain-FDR for DC are set at 5%.

4.3 Results

4.3.1 Simulations of Type I error and Power

Figure 4.2 depicted the average node-FDP and RV-level sensitivity of DATED at different layers when $M = 9$. DATED successfully controlled the node-FDR under the desired level 0.05. DATED also has increased RV-sensitivity when the number of layers increases to our suggested maximum layers (See Section 4.2.2). As the DPDs contained more and more pathogenic RVs, DATED also identifies a higher proportion of them. The performance of FL and SS P-values are similar, with DATED-SS having slightly inflated node-FDR but also slightly higher RV-sensitivity.

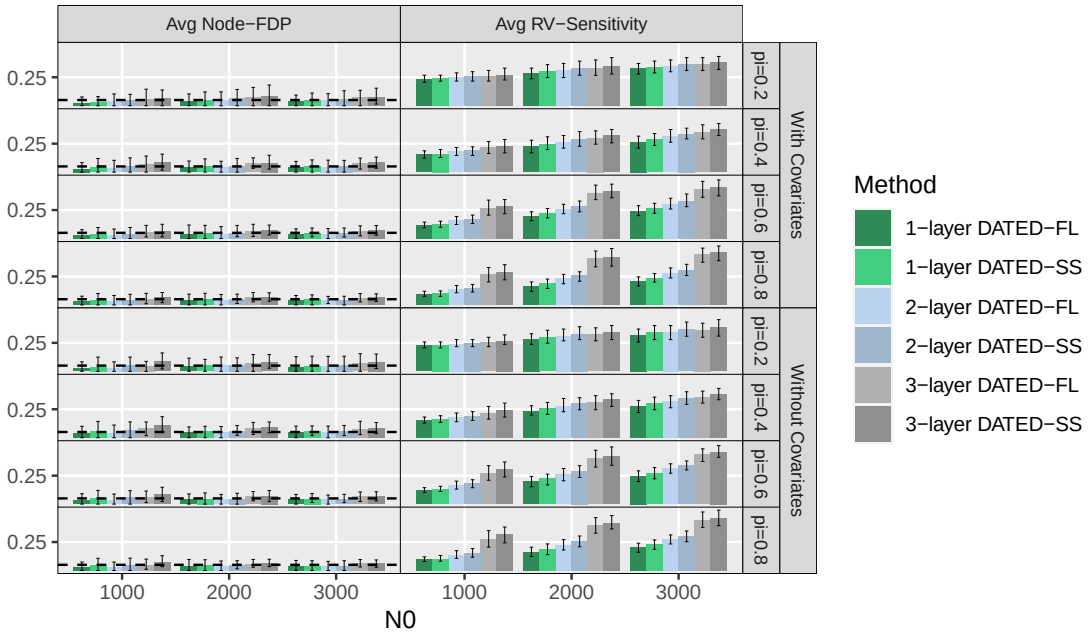


FIGURE 4.2: Layer specific performance of DATED under different simulation settings and different leaf P-values (FL and SS). The dashed horizontal lines mark where the desired FDP are. For each measure, the error bars in the bar plot are the 90% confidence intervals, which are calculated as the 5% and 95% quantiles over the 100 simulations.

We checked the performance of DATED by varying leaf sizes $M \in \{5, 7, 9, 11\}$. Figure 4.3 showed that DATED controlled the average node-FDP well under the desired level 5%. Also, DATED successfully found many regions highly enriched with pathogenic RVs: the neutral RV proportions in the DPDs are $1 - \pi$ (marked in the dashed line the middle panel of Figure 4.3); in contrast, in the DATED identified regions, the RV-FDPs are in general lower than $1 - \pi$. In addition, DATED also has robustly high RV-sensitivity. When $M \in \{5, 7, 9\}$, $L = 3$, both RV-FDP and RV-sensitivity slightly increased with M. However, when M increased to 11, the RV-sensitivity decreased because when $M = 11$, DATED only ran up to $L = 2$ layers. These results supported our suggested parameter selection procedure to identify as many disease-associated RVs as possible.

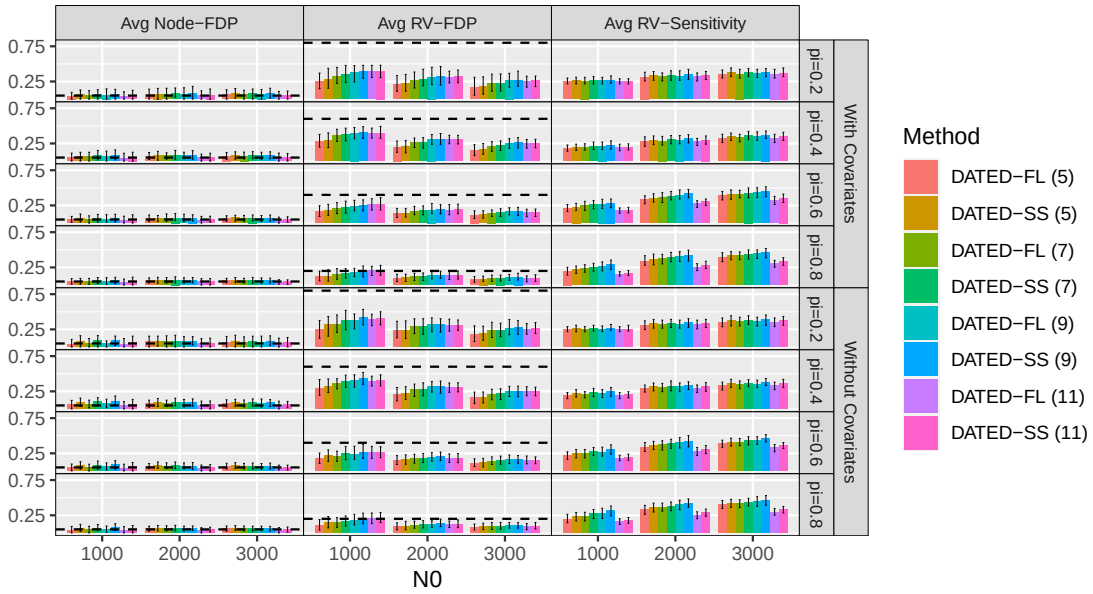


FIGURE 4.3: Performance comparison of DATED under different leaf sizes specified in the parenthesis. The comparison is performed under different simulation settings and different leaf P-values (FL and SS). The performance is measured by average GS-FDP, RV-FDP, and RV-sensitivity. The dashed horizontal line depicts the desired FDP. The error bars in the bar plot are the 90% confidence intervals, which are calculated as the 5% and 95% quantiles over the 100 simulations.

4.3.2 Comparison between DATED and alternative methods

On domain level (Figure 4.4), the DATED, DC, and SCANG have similar domain-FDP. Compared to DC, DATED consistently has higher domain-sensitivity and Domain-F1 scores. That is because the pathogenic RVs only reside in a small subregion of the domain. DC methods are suboptimal in this scenario but DATED can pinpoint the subregion by flexibly testing the regions with varying sizes. Comparing to SCANG, DATED has similar domain-sensitivity when $N_0/N_1 \in \{2, 3\}$ and higher domain-sensitivity when $N_0/N_1 = 1$. It indicates that DATED has the best domain-level performance.

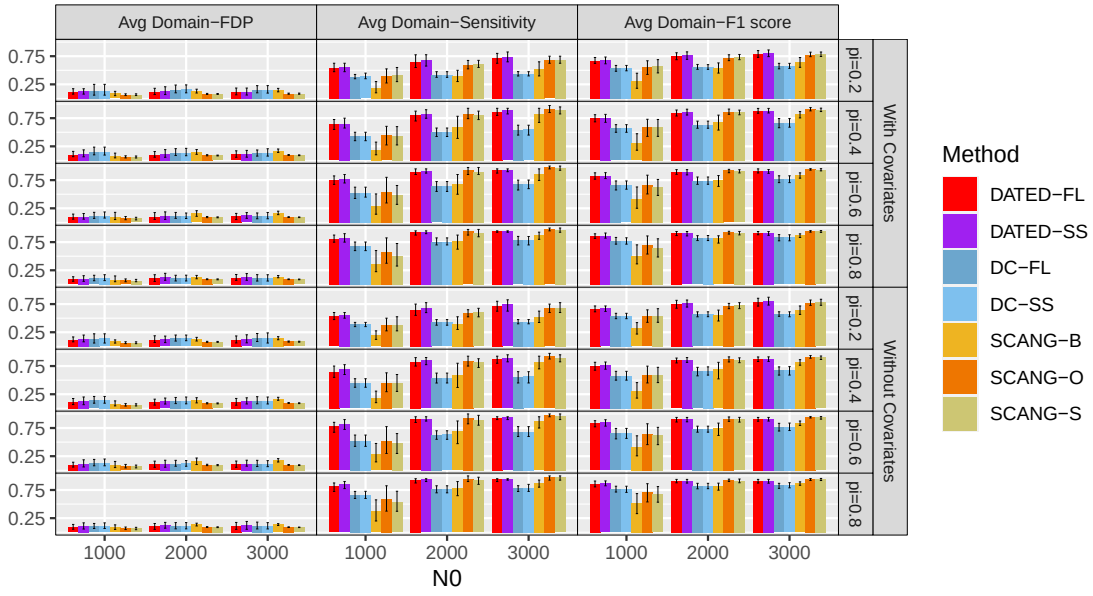


FIGURE 4.4: Performance comparison of 3-layer DATED, DC and SCANG under different simulation settings and different leaf P-values (FL and SS). The performance is measured by average domain-FDP, domain-sensitivity and domain-F1. The error bars in the bar plot are the 90% confidence intervals, which are calculated as the 5% and 95% quantiles over the 100 simulations.

Given the DC method is designed to call disease-associated domains, we only carry out the RV-level performance comparison between DATED and SCANG. Figure 4.5 summarizes this comparison. Although SCANG has a relatively larger RV-sensitivity,

its RV-FDPs are usually higher than 90%. In contrast, the mean RV-FDPs of DATED are usually less than 45% even when the proportion of neutral RV in the pathogenic region ($1 - \pi_1$) is as high as 80%.

This is perhaps not surprising to observe a high RV-sensitivity and RV-FDP in SCANG given SCANG tends to call long regions (Figure 4.6). Although a long region can easily cover many pathogenic RVs, it also covers more neutral RVs given most of the RVs are neutral. As an expense of the high RV-sensitivity, SCANG suffers severely inflated RV-FDP. A large RV-FDP indicates a low precision. Compared to SCANG, DATED output shorter region and has low RV-FDPs. By taking both RV-FDP and RV-sensitivity into account, we calculate the RV-F1 score. Under all of the scenarios, the RV-F1 scores of DATED are at least 2 times higher than the scores of SCANG. This indicates that DATED performs much better in pinpointing the pathogenic regions.

4.3.3 Application to an ALS Study

DATED versus SCANG

Compared with SCANG, the output regions from DATED tends to have shorter lengths and larger odds ratios (ORs) of RV exposure in case samples versus control samples. Specifically, the median ORs of DATED-SS and DATED-FL are both 9.6 and the median lengthes are both 14. In contrast, the median OR of SCANG range from 2.6 to 9.0, and the median length range from 28 to 41 (Figure 4.7). This indicates that DATED has higher precision compared to SCANG (Figure 4.7). From the domain level results summarized in Figure 4.8, we further notice that SCANG also fails to detect any region in the known pathogenic gene TARDBP. However, DATED successfully detect a domain in TARDBP. In addition, $^{11}/_{15}$ of the domains uniquely detected by SCANG tends to be protective (OR$\leq$ 1). Although

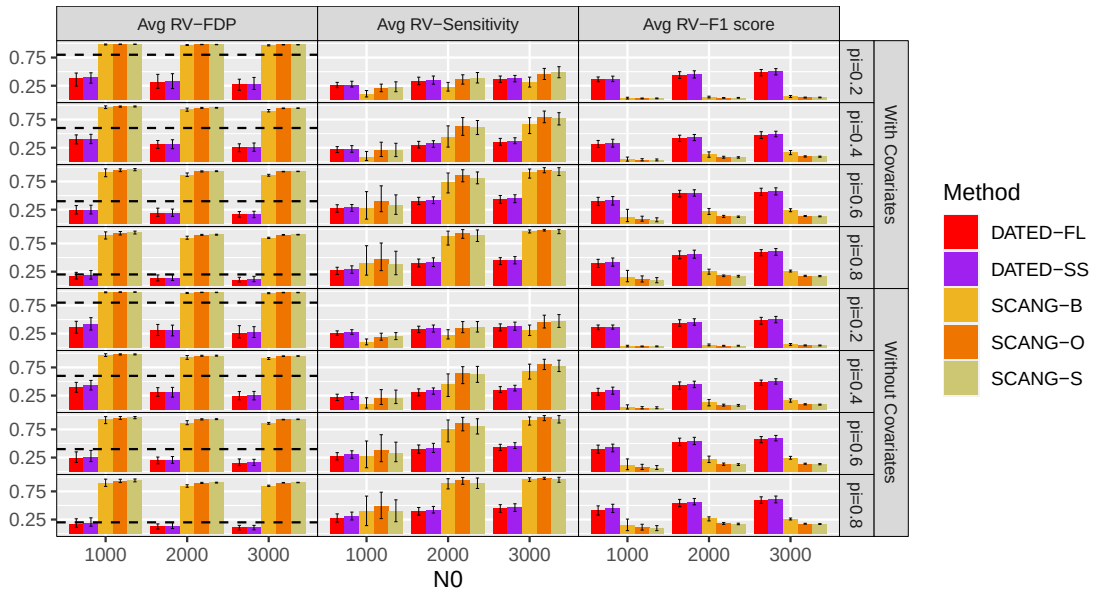


FIGURE 4.5: Performance comparison of 3-layer DATED and SCANG under different simulation settings and different leaf P-values (FL and SS). The performance is measured by average RV-FDP, RV-sensitivity, and RV-F1. The dashed horizontal line depicts the proportion of neutral RVs within the pathogenic region $(1-\pi_1)$. The error bars in the bar plot are the 90% confidence intervals, which are calculated as the 5% and 95% quantiles over the 100 simulations.

SCANG uniquely detects two potential pathogenic domains SOD1:-:-_1 (OR= ∞) and TIAM1:241264:241264_0 (OR=1.4), they have always been detected along with some potential protective domains (OR$\leq$ 1) reside in various genes (Figure 4.8). Clearly, DATED detects shorter regions and provide better targets for downstream analyses.

DATED versus DC

Given the DC method can only call disease-associated domains, we mainly compared it with DATED on the domain level. In DATED, a domain is called "detected“ if it harbors a significant RV region. Based on Figure 4.8, DATED and DC detect domain SOD1:238186:238186_0 (OR=17.3) in gene SOD1 and domain TARDBP:-:-_2 (OR=7) in gene TARDBP. Both of the genes have previously been reported to be

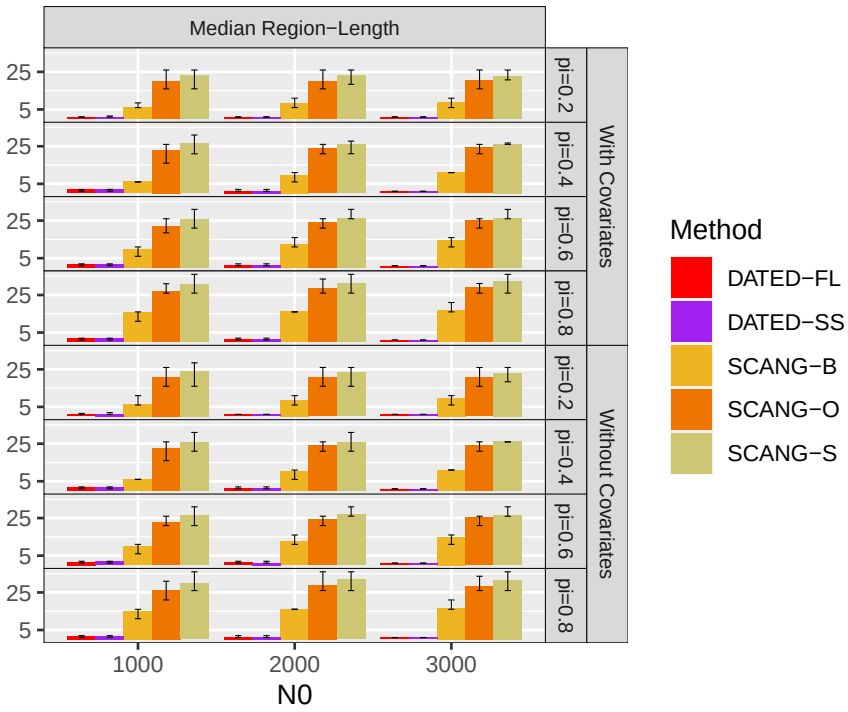


FIGURE 4.6: The median length of the detected regions. The error bars in the bar plot are the 90% confidence intervals, which are calculated as the 5% and 95% quantiles over the 100 simulations.

associated with ALS (Gelfman et al., 2019; Pesiridis et al., 2009). DATED detects an additional domain EPG5:-:-_0 (OR=1.3) reside in gene EPG5. The EPG5 is a well-known gene related to another neurodegeneration disorder named Vici Syndrome (Cullup et al., 2013; Nam et al., 2021), and one previous study find that mice with the Epg5 deficiency would exhibit key characteristics of ALS (Zhao et al., 2013). However, to the best of the authors' knowledge, there is no literature report on the association between the gene EPG5 and ALS on humans. Thus, the domain EPG5:-:-_0 is an interesting finding from DATED and is worth further investigation.

Computation time

The DC method only focuses on domain-level analysis and is thus unsurprisingly computational efficient. Here, we mainly compare the computation time between

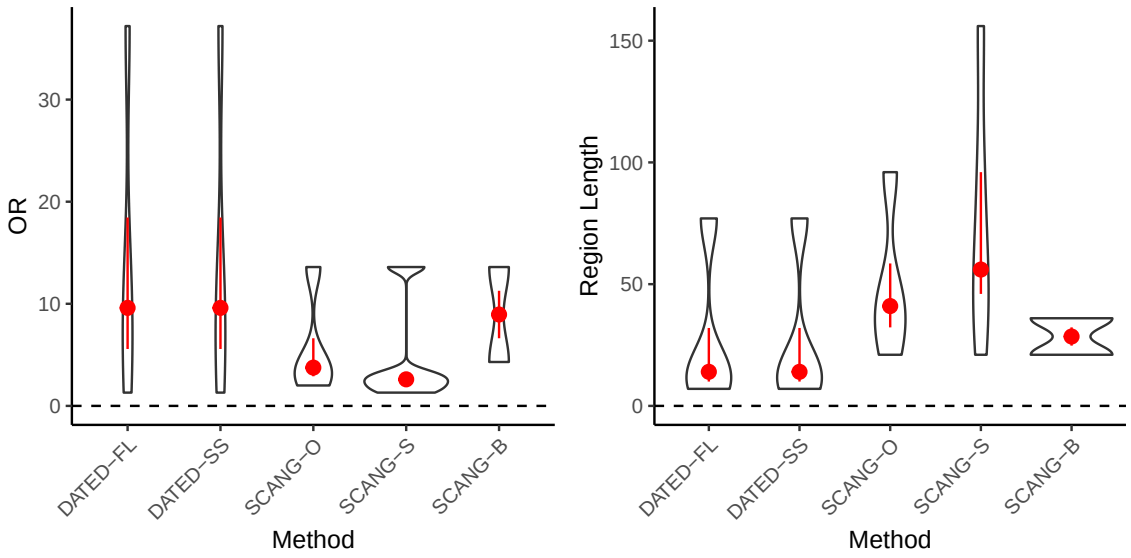


FIGURE 4.7: Violin plots depicting the distribution of detected regions' length and Odds Ratio (OR).The red dot indicates the median and the red vertical lines go from the 25th to the 75th percentile.

DATED and SCANG since both of them are varying-window methods with the potential to pinpoint the disease-associated regions. The computation time is evaluated on a 2.10 GHz Intel Xeon Gold 6252 processor with 16 Gb memory. DATED takes 1.5 minutes and 13.7 minutes by using FL and SS statistics, respectively. In contrast, the SCANG package takes 131.7 minutes to get the testing results from the three types of statistics simultaneously. Thus, on average, SCANG takes 43.9 minutes to generate results based on one statistic and is more than 3 times slower than DATED. As a varying-window method, SCANG is computationally intensive because it searches many windows with varying sizes. In contrast, DATED adopts a hierarchical aggregation strategy. The strategy avoids unnecessary consideration of many overlapping regions and thus is much faster.

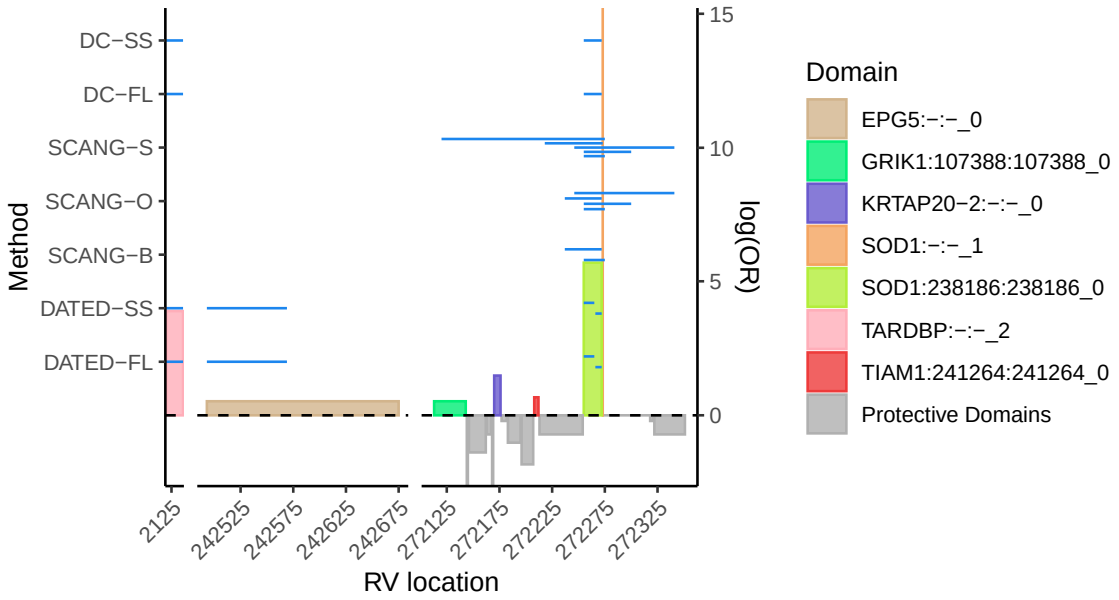


FIGURE 4.8: Genetic landscape of the RV regions detected by DATED, SCANG, and DC. The X-axis stands for the RV id ordered by their locations on the chromosomes. The RV regions detected by different methods (left y-axis) are shown in blue horizontal lines. A bar graph is used to visualize the empirical $\log(OR)$ of domains (right y-axis). Each rectangle bar represents a domain: The width of a bar represents the number of RVs that reside in the domain, and the height of a bar represents the domain level $\log(OR)$. Here, we say a domain is protective if its $\log(OR) < 0$.

4.4 Discussion

In this chapter, we developed a novel multiple testing methods named DATED to pinpoint the disease-associated rare variant regions. DATED first constructs the smallest testing unit named leaf then conducts dynamically and hierarchically testing based on the distance information. As a general testing framework, our procedure can easily accommodate different applications for different purposes.

First, DATED framework is not restricted to case-control studies with binary outcomes. As a P-value based approach, it is also applicable to studies with more complicated outcomes. For example, we can run linear regression for quantitative traits or Cox regression for the time-to-event outcome.

Second, the current leaf-level P-values are obtained by collapsing the rare variants in each leaf. The collapsing method implicitly assumes that the rare variants to be collapsed are associated with the phenotype in the same direction. This assumption is aligned with the common assumption that most of the disease-associated RVs are pathogenic. However, DATED can also easily adopt other statistics like SKAT, SKAT-O for generating the leaf-level P-values.

Last but not the least, DATED constructs and aggregates leaves based on the order of RVs. In our study, RVs are ordered by their genomic locations. We can also order RVs based on their functional annotations like the quantitative scores to measure variants' protein functions (Ng and Henikoff, 2003; Adzhubei et al., 2010), evolution conservation (Siepel et al., 2005; Pollard et al., 2010), mutation intolerance (Petrovski et al., 2013), etc. In addition, the aggregation in our study only happens within functional domains given domain-based analysis is more sensitive for pathogenic regions than gene-based analysis in ALS study (Gelfman et al., 2019). For applications with different assumptions, DATED can be easily extended to conduct aggregation within different functional regions like the gene or even the whole genome.

5

Conclusions

In this work, I participate in the development of three novel multiple testing methods that can be applied to various biomedical research. Like the typical multiple testing methods, the goals of our methods are also trying to identify a sparse amount of important features among unimportant features with controlled false discovery rates. All of our methods are designed for the data that has some ancillary information for the functional similarity of features. In Chapter 2, we introduce the DART procedure as a general framework when the features lie in a metric space. DART incorporates the functional similarity of features via constructing an aggregation tree based on the distance information and embedding multiple testing in the tree. With rigorous proof and extensive examination, we show the superiority of DART over other existing methods. In Chapter 3, we build the theoretical foundation of a method named TEAM. TEAM is designed for localizing the region where two PDFs differ. In Chapter 4, we develop a multiple testing framework named DATED for rare variant-disease association analysis. By hierarchically and dynamically aggregating rare variants and testing for their joint effect, DATED can pinpoint the disease-

associated genetic regions. A new node-level FDR is also introduced to prioritize the detection of regions with a higher concentration of disease-associated RVs.

Generally speaking, all of those three methods are hierarchical multiple testing procedures. Some other multiple testing methods also have hierarchical or graphical structures. Goeman and Finos (2012) and Meijer and Goeman (2015) developed the FWER controlling procedures on the trees and directed acyclic graphs. Dmitrienko and Tamhane (2013) developed methods testing hierarchically ordered hypotheses with applications to clinical trials and control FWER. Yekutieli (2008) considers the case when all the original hypotheses represent a node on the tree and develop a method to test those hypotheses simultaneously. Their parent-node P-values are independent from the child node P-values, which is very different from our model. Guo et al. (2018) developed a per-family error rate (PFER) and FDR controlling procedure for hypotheses with a DAG structure. Li et al. (2020) developed a bottom-up multiple testing approach embedded in the aggregation tree.

Although those existing hierarchical multiple testing procedures share some similarities with our methods, their settings and focuses are different. First, the hierarchical structures of our methods are not given but constructed from the distance matrix or localizing structure of data (e.g. genomics locations). Second, our testing nodes depend on the previous rejection path. One of our main contributions is to introduce dynamic hypotheses in the hierarchical testing framework and to develop new methods that asymptotically control FDR under this case. Third, many existing papers focus on prefixed-node-level FDR or FWER control, while DART and TEAM focus on feature-level FDR control. The feature-level FDR control can be applied to a wide range of application contexts with the purpose of feature selection. DATED also introduce a novel FDR weighted by the proportion of neutral RVs. Last but not the least, many existing papers assume the feature or node P-values are uniformly

distributed (Unif$(0,1)$) under the null. In practice, most feature P-values are derived from asymptotic tests, such as the Wald tests, the score tests, the likelihood ratio tests, etc., and thus their null distributions are not exactly Unif$(0,1)$. Although under most scenarios, they weakly converge to Unif$(0,1)$, under high-dimensional settings (with much more features than sample size), the distribution deviation may inflate the FDR. In my dissertation, we are focusing on much more complicated null distributions for the working P-values and thus our methods would have broader application.

Appendix A

Appendix for Chapter 2

A.1 Algorithm Pseudo Codes

A.1.1 Stage I: Transform the distance matrix into an aggregation tree

To obtain such an aggregation tree, we develop Algorithm 3 with remarks listed below.

Remark A.1. On layer 1, we set up each node as a single feature node. All these nodes have empty children sets.

Remark A.2. On layer ℓ, we aggregate nodes from layer $\ell - 1$ to form new nodes on this layer.

Remark A.3. At the beginning of layer ℓ, $A^{(\ell)}$ is set as the empty set, and it will be updated during the aggregation process. $\tilde{\mathcal{A}}$ is the candidate node set with all the nodes that can possibly be aggregated. It may contain the layer $\ell - 1$'s nodes that have not be aggregated yet and layer ℓ's nodes that have already been aggregated but can possibly be further aggregated. The layer ℓ distance between $A_1, A_2 \in \tilde{\mathcal{A}}$ is denoted by $\text{dist}^{(\ell)}(A_1, A_2)$. We set it equals to $\text{dist}(A_1, A_2)$, which is defined in

Algorithm 3: DART Stage I. Transform the distance matrix into an aggregation tree.

Data: distance matrix $\boldsymbol{D} = (d_{ij})_{m\times m}$, the maximum layer L, the maximum children number M, the maximum distance threshold $g^{(2)}, \ldots, g^{(L)}$.

Result: an aggregation tree $\mathcal{T}_L = \{\mathcal{A}^{(\ell)} : \ell \in \{1, \ldots, L\}\}$ and $\mathcal{C}(A)$ for all $A \in \mathcal{A}^{(\ell)}$ and all $\ell \in \{1, \ldots, L\}$.

$\ell = 1$, $\mathcal{A}^{(1)} = \{\{1\}, \ldots, \{m\}\}$,
for $A \in \mathcal{A}^{(1)}$ **do** $\mathcal{C}(A) = \emptyset$ *// Remark A.1*
for $\ell \in \{2, \ldots, L\}$ **do** *// Remark A.2*
  $\mathcal{A}^{(\ell)} = \emptyset$, $\tilde{\mathcal{A}} = \mathcal{A}^{(\ell-1)}$, $\text{dist}^{(\ell)}(A_1, A_2) = \text{dist}(A_1, A_2), \forall A_1, A_2 \in \tilde{\mathcal{A}}$
  // Remark A.3
  while $|\tilde{\mathcal{A}} \setminus \mathcal{A}^{(\ell)}| > 0$ **do**
    $(\breve{A}_1, \breve{A}_2) = \underset{A_1\in\tilde{\mathcal{A}},\ A_2\in\tilde{\mathcal{A}}\setminus\{A_1\}}{\arg\min} \text{dist}^{(\ell)}(A_1, A_2)$ *// Remark A.4*
    if $\text{dist}^{(\ell)}(\breve{A}_1, \breve{A}_2) > g^{(\ell)}$ **then** *// Remark A.5*
      for $A \in \tilde{\mathcal{A}} \setminus \mathcal{A}^{(\ell)}$ **do** $\mathcal{C}(A) = A$, $\mathcal{A}^{(\ell)} = \mathcal{A}^{(\ell)} \cup \{A\}$, $\tilde{\mathcal{A}} = \tilde{\mathcal{A}} \setminus \{A\}$
    else
      $\breve{A} = \breve{A}_1 \cup \breve{A}_2$
      for $i \in \{1, 2\}$ **do** *// Remark A.6*
        if $\breve{A}_i \in \mathcal{A}^{(\ell)}$ **then** $\mathcal{C}_i = \mathcal{C}(\breve{A}_i)$ **else** $\mathcal{C}_i = \{\breve{A}_i\}$
      $\mathcal{C}(\breve{A}) = \mathcal{C}_1 \cup \mathcal{C}_2$
      if $|\mathcal{C}(\breve{A})| < M$ **then**
        $\mathcal{A}^{(\ell)} = \mathcal{A}^{(\ell)} \cup \{\breve{A}\} \setminus \{\breve{A}_1, \breve{A}_2\}$, $\tilde{\mathcal{A}} = \tilde{\mathcal{A}} \cup \{\breve{A}\} \setminus \{\breve{A}_1, \breve{A}_2\}$
      else if $|\mathcal{C}(\breve{A})| = M$ **then**
        $\mathcal{A}^{(\ell)} = \mathcal{A}^{(\ell)} \cup \{\breve{A}\} \setminus \{\breve{A}_1, \breve{A}_2\}$, $\tilde{\mathcal{A}} = \tilde{\mathcal{A}} \setminus \{\breve{A}_1, \breve{A}_2\}$
      else
        $\text{dist}^{(\ell)}(\breve{A}_1, \breve{A}_2) = +\infty$ *// Remark A.7*

section 2.2.2.

Remark A.4. We use the greedy algorithm to select the closest two nodes $\breve{A}_1$ and $\breve{A}_2$ from the current candidate node set $\tilde{\mathcal{A}}$. If there exists a tie, we select the first node pair that reaches the minimal distance. For example, in Figure 2.1b, at the beginning of layer 2, $\text{dist}(\{1\}, \{2\}) = 2$ reaches the minimal distance among all node pairs on layer 1, so they will be selected to be further considered for aggregation.

Remark A.5. We check if $\text{dist}(\breve{A}_1, \breve{A}_2) > g^{(\ell)}$. If yes, the remaining candidate nodes

are too far away from each other and will not be further aggregated. Then the remaining child nodes on layer $\ell - 1$ will be kept on layer ℓ, and the aggregation on layer ℓ ends. If not, $\breve{A}_1$ and $\breve{A}_2$ will be further considered for aggregation.

Remark A.6. We define the new node $\breve{A} = \breve{A}_1 \cup \breve{A}_2$. $\mathcal{C}(\breve{A})$ depends on the identity of $\breve{A}_1$ and $\breve{A}_2$: if $\breve{A}_i$ is a candidate child on layer $\ell - 1$, then itself will be included in $\mathcal{C}(\breve{A})$; otherwise, $\breve{A}_i$'s children will be included in $\mathcal{C}(\breve{A})$.

Remark A.7. We check the number of children of $\breve{A}$. If $|\mathcal{C}(\breve{A})| < M$, we add $\breve{A}$ to $\mathcal{A}^{(\ell)}$ and remove $\breve{A}_1$ and $\breve{A}_2$, and change $\tilde{\mathcal{A}}$ correspondingly. If $|\mathcal{C}(\breve{A})| = M$, we change $\mathcal{A}^{(\ell)}$ in the same way when $|\mathcal{C}(\breve{A})| < M$, and remove $\breve{A}_1$ and $\breve{A}_2$ from $\tilde{\mathcal{A}}$ to prevent them being selected again. This step also guarantees the number of children of a node $A \in \tilde{\mathcal{A}}$ is always smaller than M. If $|\mathcal{C}(\breve{A})| > M$, we just reset the layer ℓ distance between $\breve{A}_1$ and $\breve{A}_2$ to be $+\infty$, so that they will never be aggregated on layer ℓ, but still have chance to aggregate with other nodes in $\tilde{\mathcal{A}}$.

A.1.2 Stage II: Embed multiple testing in the tree

Algorithm 4: DART Stage II. Embed multiple testing in the tree.

Data: Tree $\mathcal{T}_L = \{\mathcal{A}^{(i)} : i = 1, \ldots, L\}$, P-values $(T_1, \ldots, T_m)$, and FDR level α.

Result: The set of rejected features R_{feat}.

Set $\hat{t}^{(1)}$ as in (2.6), $R_{\text{feat}} = \{i : T_i \leq \hat{t}^{(1)}\}$ *// Multiple testing on layer 1.*

for $\ell \in \{2, \ldots, L\}$ **do** *// Testing recursively on higher layers*

  $k = 0$, $T =$NULL

  for $S_k \in \mathcal{B}^{(\ell)}$ **do**

    $k = k + 1$, $X_{S_k} = \sum_{j \in S_k} \bar{\Phi}^{-1}(T_j)/\sqrt{|S_k|}$

    $T = (T, \bar{\Phi}(X_{S_k}))'$ *// Append the new working P-value at the end*

  Set $\hat{t}^{(\ell)}$ as in (2.8), $R^{(\ell)}_{\text{node}} = \{S_{k'} : T_{k'} < \hat{t}^{(\ell)}\}$, $R_{\text{feat}} = R_{\text{feat}} \cup \{\cup_{S \in R^{(\ell)}_{\text{node}}} S\}$

A.1.3 Tuning parameter selection for applying DART on simulated data

The section 2.3 introduces the tuning parameter selection for the aggregation tree construction. Based on it, the tuning parameter for our numerical study is selected as follow:

- If $(n, m) = (90, 100)$: Based on recommendation in section 2.3, we choose $M = 3$ and construct a $L = \lceil \log_M 100 - \log_M 30 \rceil = 2$ layers aggregation tree. We use Algorithm 5 to construct a dynamic set G and search the value $g^{(2)}$. Table A.1 (1) tracks the number of non-single-child nodes $|\tilde{A}^{(2)}(g)|$ based on different values of $g \in G$. By applying the algorithm,

$$g^{(2)} = 26/\sqrt{n \log m \log\log m}$$

- If $(n, m) = (300, 1000)$: Similar to the previous case, we choose $M = 3$ and construct a $L = \lceil \log_M 1000 - \log_M 30 \rceil = 4$ layers aggregation tree. Based on Algorithm 5, we have Table A.1 which tracks the number of non-single-child nodes on each layer, and,

$$g^{(2)} = \frac{8}{\sqrt{n \log m \log\log m}}, g^{(3)} = \frac{22}{\sqrt{n \log m \log\log m}}, g^{(4)} = \frac{56}{\sqrt{n \log m \log\log m}}$$

Algorithm 5: $g^{(\ell)}$ Selection algorithm.

Data: Distance Matrix $D = (d_{ij})_{m\times m}$, Sample size n, number of features m, the maximum children number M, the maximum layer L

Result: $g^{(2)}, ..., g^{(L)}$.

// set searching upper bound $d_{\max}$ and step-size $s_{n,m}$

Let $d_{\max} = \max_{j\in\Omega} \min_{i\in\{i:i\neq j\}} d_{ij}$; $s_{n,m} = 2/\sqrt{n\log(m)\log\log(m)}$;

for $\ell = 2, ..., L$ **do**

// on layer ℓ, search $g^{(\ell)}$ from $(g^{(\ell-1)}, d_{\max}]$, $g^{(1)} = 0$

Let M_g =NULL; e_g=1; G =NULL; $g = g^{(\ell-1)} + s_{n,m}$;

while $g \leq (2M^{L-2} - 1)d_{\max}$ *and* $e_g < 10$ **do**

// stop searching process if the value g exceed the searching upper bound or the $|\tilde{A}^{(\ell)}(g)|$ does not increase for past 10 candidate values g.

// stop searching process if the value g exceed the searching upper bound.

Use Algorithm 3 to Construct an ℓ layers aggregation tree $\mathcal{T}_\ell = \{\mathcal{A}^{(\ell')} : \ell' = 1, ..., \ell\}$ with maximum children number M, and $(g^{(1)}, ..., g^{(\ell-1)}, g)$;

Set $\tilde{A}^{(\ell)}(g) = \{A : A \in \mathcal{A}^{(\ell)}(g),\ |\mathcal{C}(A)| \geq 2\}$; **if** $m_g \geq |\tilde{A}^{(\ell)}(g)|$ **then**

$e_g = e_g + 1$;

else

e_g=1;

$G = (G, g)$; $M_g = (M_g, |\tilde{A}^{(\ell)}(g)|)$; $m_g = |\tilde{A}^{(\ell)}(g)|$;

$g = g + s_{n,m}$;

$g^{(\ell)} = \min\{\arg\max_{g\in G} M_g\}$;

A.2 Additional numerical results for assessing impact of the parameter M

In this section, we numerically investigate the impact of the choice of M by comparing the numerical results when $M = 3$ and $M = \infty$. When $M = 3$, the tunning parameters are same to the parameters in 2.2.3. When $M = \infty$, in order to have a relatively fair comparison, we set the same total layer L as the value in 2.2.3. The selection procedure of $g^{(\ell)}$ is similar to 2.2.3. Based on the Algorithm 5,we have

- If $(n, m) = (90, 100)$: we set $g^{(2)} = \frac{16}{\sqrt{n\log m\log\log m}}$

Table A.1: The number of non-single-child nodes based on value $g \in G$ when $M = 3$. For simplicity purpose, the value g is represented by its nominator: $g' = g \times \sqrt{n \log m \log\log m}$. The selected g' and its correspnding $|\tilde{\mathcal{A}}^{(2)}(g)|$ is highlighted in bold.

(1) $(n, m) = (90, 100)$:

Layer 2	g'	2	4	6	8	10	12	14	16	18	20	22	24	**26**	28	30	32	...
	$\lvert\tilde{\mathcal{A}}^{(3)}(g)\rvert$	5	10	17	22	29	31	31	39	40	40	40	40	**41**	41	41	41	...

(2) $(n, m) = (300, 1000)$:

Layer 2	g'	2	4	6	**8**	10	12	14	16	...								
	$\lvert\tilde{\mathcal{A}}^{(2)}(g)\rvert$	49	149	245	**293**	293	293	293	293	...								
Layer 3	g'	10	12	14	16	18	20	**22**	24	26	28	30	...					
	$\lvert\tilde{\mathcal{A}}^{(3)}(g)\rvert$	103	154	191	221	230	239	**241**	241	241	241	241	...					
Layer 4	g'	...	38	40	42	44	46	48	50	52	54	**56**	58	60	62	64	...	
	$\lvert\tilde{\mathcal{A}}^{(4)}(g)\rvert$	...	116	118	119	120	120	120	120	120	120	**121**	121	121	121	121	...	

- If $(n, m) = (300, 1000)$: we set $g^{(2)} = \frac{12}{\sqrt{n \log m \log\log m}}$, $g^{(3)} = \frac{26}{\sqrt{n \log m \log\log m}}$ and $g^{(4)} = \frac{44}{\sqrt{n \log m \log\log m}}$.

Figure A.1 compares the performance between two different M values under SE1-SE5. We only compare the performance on the top layer of the aggregation tree. Based on the figure, our method is still valid with FDR control when $M = \infty$.

Table A.2: The number of non-single-child nodes based on value $g \in G^{(\ell)}$ when $M = \infty$. For simplicity purpose, the value g is represented by its nominator: $g' = g \times \sqrt{n \log m \log\log m}$. The selected g' and its correspnding $|\tilde{\mathcal{A}}^{(\ell)}(g)|$ is highlighted in bold.

(1) $(n, m) = (90, 100)$:

Layer 2	g'	2	4	6	8	10	12	14	**16**	18	20	22	24	...
	$\lvert\tilde{\mathcal{A}}^{(2)}(g)\rvert$	5	10	17	22	29	30	30	**35**	32	31	29	28	...

(2) $(n, m) = (300, 1000)$:

Layer 2	g'	2	4	6	8	10	**12**	14	16	18	20	...			
	$\lvert\tilde{\mathcal{A}}^{(2)}(g)\rvert$	49	149	239	291	300	**303**	295	288	263	245	...			
Layer 3	g'	14	16	18	20	22	24	**26**	28	30	32	34	...		
	$\lvert\tilde{\mathcal{A}}^{(3)}(g)\rvert$	53	101	130	148	163	167	**169**	166	159	152	144	...		
Layer 4	g'	28	30	32	34	36	38	40	42	**44**	46	48	50	52	...
	$\lvert\tilde{\mathcal{A}}^{(4)}(g)\rvert$	17	33	47	56	66	70	74	76	**80**	79	79	79	77	...

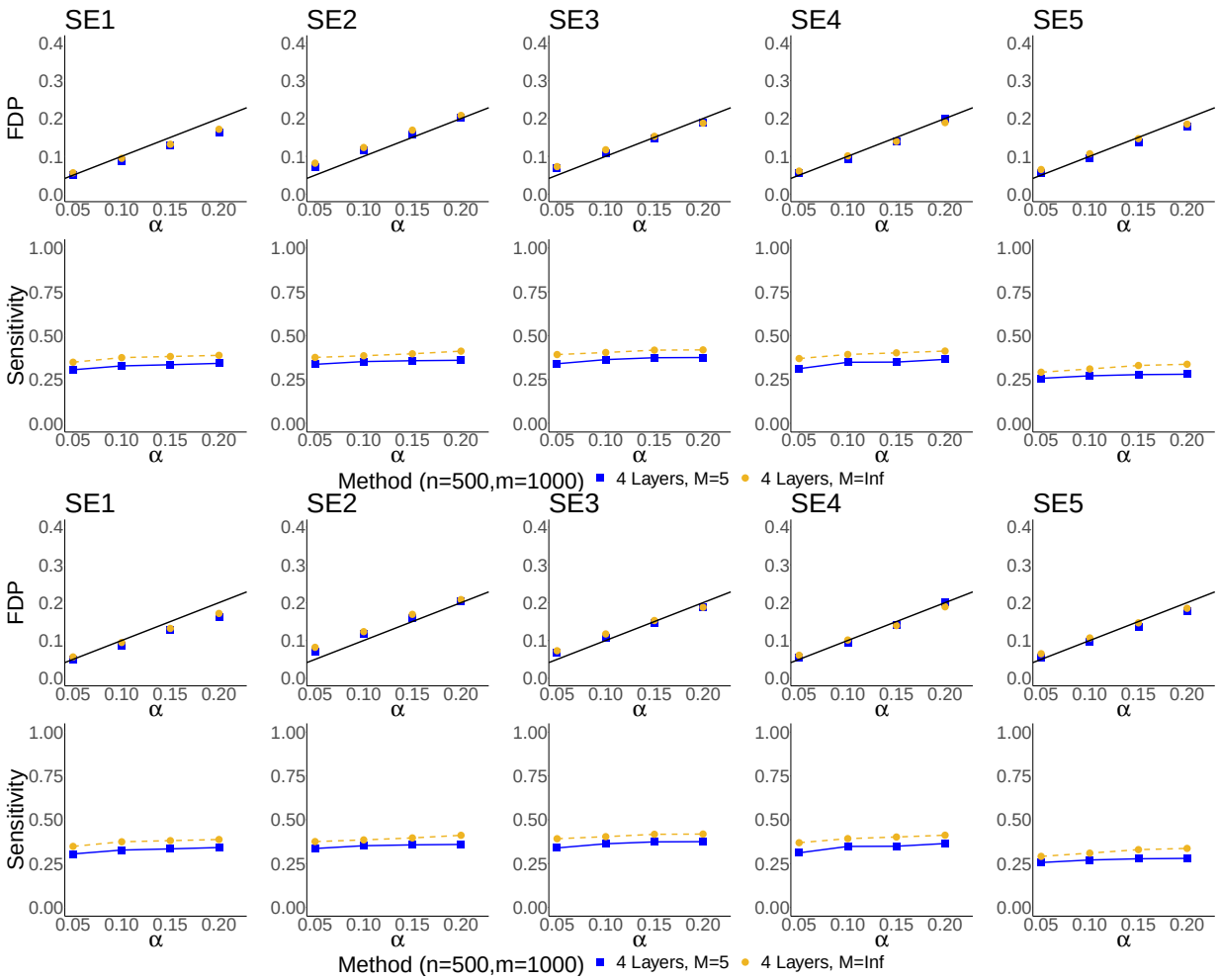


FIGURE A.1: *Additional simulation results for setting SE1-SE5. The first two rows represent the results in the setting* $(n, m) = (90, 100)$*, and the second two rows represent the results in the setting* $(n, m) = (300, 1000)$.

A.3 Proof of the Lemmas

Proof of Lemma 2.1. Let $X_i'^* = \frac{\boldsymbol{q}^T\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}}{s\sqrt{\boldsymbol{q}^T(\boldsymbol{\mathcal{W}}^T\boldsymbol{\mathcal{W}})^{-1}\boldsymbol{q}}}$ and $X'_{o,i} = \frac{\boldsymbol{q}^T\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}}{\sigma\sqrt{\boldsymbol{q}^T(\boldsymbol{\mathcal{W}}^T\boldsymbol{\mathcal{W}})^{-1}\boldsymbol{q}}}$, we have $X'_{o,i} \sim N(\eta_i, 1)$ with $\eta_i = \frac{\boldsymbol{q}^T\boldsymbol{\theta}_i}{\sigma\sqrt{\boldsymbol{q}^T(\boldsymbol{\mathcal{W}}^T\boldsymbol{\mathcal{W}})^{-1}\boldsymbol{q}}}$.

To show the statistics T_i is asymptotically oracle, it is suffice to show:

$$P(|\Phi^{-1}(T_i) - \Phi^{-1}(\tilde{T}_i)| > (\log m)^{-2.5}) = o((\log m)^{-1})$$

Given

$$|\Phi^{-1}(T_i) - \Phi^{-1}(\tilde{T}_i)| \leq \sup_{x \geq 0} \frac{\phi(x)}{\phi[\Phi^{-1}(2\Phi(-x))]} ||X_i^{'*}| - |X'_{o,i}||$$

$$\leq |X_i^{'*} - X'_{o,i}|$$

$$= |X'_{o,i}(\sigma/s - 1)|$$

and based on condition 2.1,

$$P\bigg(|X'_{o,i}(\sigma/s - 1)| > (\log m)^{-2.5}\bigg) \leq P(|X'_{o,i}| > \sqrt{\log m}) + P(|\sigma/s - 1| > (\log m)^{-3})$$

It is suffice to show

$$P(|\sigma/s - 1| > (\log m)^{-3}) = o((\log m)^{-1}) \tag{A.1}$$

Let $Y_1, ..., Y_{n-p_0-1} \overset{iid}{\sim} \mathcal{X}^2(1)$ and $Y = \sum_{k=1}^{n-p_0-1}(Y_k - 1)/\sqrt{2(n-p_0-1)}$, we have $Y/\sqrt{n-p_0-1} \sim \mathcal{X}^2(n-p_0-1)/(n-p_0-1) - 1$. Since $s^2/\sigma^2 \sim \mathcal{X}^2(n-p_0-1)/(n-p_0-1)$, based on Lemma 6.1 in Liu et al. (2013),

$$P(|s/\sigma - 1| > (\log m)^{-3}) \leq P(|s^2/\sigma^2 - 1| > (\log m)^{-3}) = o((\log m)^{-1})$$

Thus, after trivial calculation, the equation (A.1) holds. □

Proof of Lemma 2.2. (1) Define $\tilde{X}_i = \bar{\Phi}(\tilde{T}_i)$, For $k \in \{1, ..., c_0\}$, let $q_0 \geq \epsilon(m)$. Also define $b_{1,k}(q_0)$, c_1,...,c_k be the value s.t. $P(\sum_{j=1}^{k} \tilde{X}_j > b_{1,k}(q_0)) = q_0\big[\epsilon'(m)\big]^{(c_0-k)/c_0}$, and $P(\tilde{X}_1 > c_1) = ... = P(\tilde{X}_k > c_k) = \epsilon(m)\epsilon'(m)$, respectively. For simplicity sake, we use $b_{1,k}$ to present $b_{1,k}(q_0)$.

Based on the definition, we have

$$b_{1,k} < \sum_{j=1}^{k} c_j$$

Thus, when $k = 2$,

$$\begin{aligned}
&P(\hat{X}_1 + \hat{X}_2 > b_{1,2}) \\
=&P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_1 > b_{1,2} - c_2, \hat{X}_2 > b_{1,2} - c_1) \\
&+ P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_1 < b_{1,2} - c_2) + P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_2 < b_{1,2} - c_1) \\
=&P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, c_1 > \hat{X}_1 > b_{1,2} - c_2) + P(\hat{X}_1 > c_1, \hat{X}_2 > b_{1,2} - c_1) \\
&+ P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_1 < b_{1,2} - c_2) + P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, \hat{X}_2 < b_{1,2} - c_1)
\end{aligned}$$

Based on construction, the last three terms always smaller than $\epsilon(m)\epsilon'(m)(1+\delta_4(m))$ for $\delta_4(m) := \max_{i\in\Omega} \sup_{p\in\mathcal{P}_i'} \left| P(\hat{T}_i < p)/P(\tilde{T}_i < p) - 1 \right| \to 0$, and accordingly, we have

$$\begin{aligned}
&P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, c_1 > \hat{X}_1 > b_{1,2} - c_2) + P(\hat{X}_1 > c_1, \hat{X}_2 > b_{1,2} - c_1) \\
\leq&[P(\hat{X}_1 + \tilde{X}_2 > b_{1,2}, c_1 > \hat{X}_1 > b_{1,2} - c_2) + P(\hat{X}_1 > c_1, \tilde{X}_2 > b_{1,2} - c_1)](1 + \delta_4(m)) \\
\leq&[P(\hat{X}_1 + \tilde{X}_2 > b_{1,2}, \hat{X}_1 > b_{1,2} - c_2, \tilde{X}_2 > b_{1,2} - c_1)](1 + \delta_4(m)) \\
\leq&P(\tilde{X}_1 + \tilde{X}_2 > b_{1,2}, \tilde{X}_1 > b_{1,2} - c_2, \tilde{X}_2 > b_{1,2} - c_1)(1 + \delta_4(m))^2
\end{aligned}$$

Based on similar arguments, we can also have

$$\begin{aligned}
&P(\hat{X}_1 + \hat{X}_2 > b_{1,2}, c_1 > \hat{X}_1 > b_{1,2} - c_2) + P(\hat{X}_1 > c_1, \hat{X}_2 > b_{1,2} - c_1) \\
\geq&P(\tilde{X}_1 + \tilde{X}_2 > b_{1,2}, \tilde{X}_1 > b_{1,2} - c_2, \tilde{X}_2 > b_{1,2} - c_1)(1 - \delta_4(m))^2
\end{aligned}$$

Thus,

$$\sup_{q_0 \geq \epsilon(m)\left[\epsilon'(m)\right]^{\frac{c_0-2}{c_0}}} \left| \frac{P(\hat{X}_1 + \hat{X}_2 > b_{1,2})}{P(\tilde{X}_1 + \tilde{X}_2 > b_{1,2})} - 1 \right| \to 0$$

Similarly, if $\sup_{q_0 \geq \epsilon(m)\left[\epsilon'(m)\right]^{\frac{c_0-k}{c_0}}} \left| \frac{P(\sum_{j=1}^{k} \hat{X}_j > b_{1,k})}{P(\sum_{j=1}^{k} \tilde{X}_j > b_{1,k})} - 1 \right| \to 0$, we can have

$$\sup_{q_0 \geq \epsilon(m)\left[\epsilon'(m)\right]^{\frac{c_0-k-1}{c_0}}} \left| \frac{P(\sum_{j=1}^{k+1} \hat{X}_j > b_{1,k+1})}{P(\sum_{j=1}^{k+1} \tilde{X}_j > b_{1,k+1})} - 1 \right| \to 0$$

Thus, we can get (1). In addition, based on the similar arguments, we can get (2).

□

Proof of Lemma 2.3. (1) Let $Z_1', ..., Z_K' \overset{iid}{\sim} N(0,1)$, with $2 \leq K < M^{L-1}$. Define the set $\mathfrak{M} = \{\mathcal{M}_1 \subset \{1, ..., m\} : 1 \leq |\mathcal{M}_1| \leq K-1\}$. It is suffice to show:

$$\lim_{m\to\infty} \sup_{\mathcal{M}_1 \in \mathfrak{M}} \sup_{\substack{c_1 \in [\beta_0, \gamma_m] \\ c_2 \in [0, \gamma_m]}} \frac{\mathsf{P}(\frac{1}{\sqrt{K}} \sum_{i=1}^K Z_i' > c_2, \frac{1}{\sqrt{|\mathcal{M}_1|}} \sum_{j \in \mathcal{M}_1} Z_j' > c_1)}{\mathsf{P}(\frac{1}{\sqrt{K}} \sum_{i=1}^K Z_i' > c_2)} = 0$$

Here, $\beta_0 = \sqrt{2b(1-r_1)\log m + b(1-r_1)\log\log\log m}$, with

$$b = \frac{\frac{2M^{L-1}+1}{M^{L-1}+1} - r_1}{2(1-r_1)} \in \left(\frac{M^{L-1}}{(M^{L-1}+1)(1-r_1)}, 1 \right).$$

For simplification, let $k_1 = |\mathcal{M}_1|$. For Z_1 and $Z_2 \overset{iid}{\sim} \mathrm{N}(0,1)$, define

$$\mathcal{D}_m = \left\{ c_2 \in (0, \gamma_m) : \frac{d}{dc_2} \frac{P\big(\sqrt{\frac{k_1}{K}} Z_1 + \sqrt{\frac{K-k_1}{K}} Z_2 > c_2, Z_1 > \beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}} Z_1 + \sqrt{\frac{K-k_1}{K}} Z_2 > c_2\big)} = 0 \right\}$$

, then

$$\sup_{\substack{c_1 \in [\beta_0, \gamma_m] \\ c_2 \in [0, \gamma_m]}} \frac{P(\frac{1}{\sqrt{K}} \sum_{i=1}^K Z_i' > c_2, \frac{1}{\sqrt{|\mathcal{M}_1|}} \sum_{j \in \mathcal{M}_1} Z_j' > c_1)}{P(\frac{1}{\sqrt{K}} \sum_{i=1}^K Z_i' > c_2)}$$

$$\leq 2 \sup_{c_2 \in [0, \gamma_m]} \frac{P\big(\sqrt{\frac{k_1}{K}} Z_1 + \sqrt{\frac{K-k_1}{K}} Z_2 > c_2, Z_1 > \beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}} Z_1 + \sqrt{\frac{K-k_1}{K}} Z_2 > c_2\big)}$$

$$\leq 2 \max \left\{ \max_{c_2 = 0 \text{ or } \gamma_m} \frac{P\big(\sqrt{\frac{k_1}{K}} Z_1 + \sqrt{\frac{K-k_1}{K}} Z_2 > c_2, Z_1 > \beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}} Z_1 + \sqrt{\frac{K-k_1}{K}} Z_2 > c_2\big)}, \right.$$

$$\left. \sup_{c_2 \in \mathcal{D}_m} \frac{P\big(\sqrt{\frac{k_1}{K}} Z_1 + \sqrt{\frac{K-k_1}{K}} Z_2 > c_2, Z_1 > \beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}} Z_1 + \sqrt{\frac{K-k_1}{K}} Z_2 > c_2\big)} \right\}$$

(i). When $c_2 = 0$,

$$\lim_{m\to\infty} \frac{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1 > \beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)} = \lim_{m\to\infty} 2P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1 > \beta_0\big) = 0$$

(ii). When $c_2 = \gamma_m$, $c_2/\beta_0 = \sqrt{\frac{1}{b(1-r_1)}}$,

$$\lim_{m\to\infty} \frac{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1 > \beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)} = \lim_{\beta_0\to\infty} \frac{\int_{\beta_0}^{\infty}\int_{S\sqrt{\frac{K}{K-k_1}}\beta_0 - \sqrt{\frac{k_1}{K-k_1}}z_1}^{\infty} \phi(z_1)\phi(z_2)dz_2dz_1}{\int_{S\beta_0}^{\infty}\phi(z)dz}$$

$$\leq C \lim_{\beta_0\to\infty} \frac{\int_{S\sqrt{\frac{K}{K-k_1}}\beta_0 - \sqrt{\frac{k_1}{K-k_1}}\beta_0}^{\infty} \phi(\beta_0)\phi(z)dz + \int_{\beta_0}^{\infty}\phi(z)\phi(S\sqrt{\frac{K}{K-k_1}}\beta_0 - \sqrt{\frac{k_1}{K-k_1}}z)dz}{\phi(S\beta_0)} \text{ (L'Hopital's rule)}$$

$$\leq C \lim_{\beta_0\to\infty} \left[\exp\Big\{-\frac{\beta_0^2}{2}\Big(S\sqrt{\frac{k_1}{K-k_1}} - \sqrt{\frac{K}{K-k_1}}\Big)^2\Big\} + \int_{\beta_0}^{\infty}\exp\Big\{-\frac{1}{2}\Big(\sqrt{\frac{K}{K-k_1}}z - S\sqrt{\frac{k_1}{K-k_1}}\beta_0\Big)^2\Big\}dz\right] = 0,$$

Where $S = \sqrt{\frac{1}{b(1-r_1)}}$

(iii). When $c_2 \in \mathcal{D}_m$, given

$$\begin{aligned}
0 =& \frac{d}{dc_2}\frac{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1 > \beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)} \\
=& \frac{1}{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)^2}\times \\
& \Big\{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1 > \beta_0\big) \\
& - P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1 > \beta_0\big)\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)\Big\}
\end{aligned}$$

We have

$$\frac{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1 > \beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)} = \frac{\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2, Z_1 > \beta_0\big)}{\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1 + \sqrt{\frac{K-k_1}{K}}Z_2 > c_2\big)}$$

Therefore,

$$\sup_{c_2\in\mathcal{D}_m}\frac{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0\big)}{P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2\big)}$$

$$=\sup_{c_2\in\mathcal{D}_m}\frac{\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0\big)}{\frac{d}{dc_2}P\big(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2\big)}$$

$$=\sup_{c_2\in\mathcal{D}_m}C\int_{\beta_0}^{\infty}\exp\Big\{-\frac{1}{2}\Big(\sqrt{\frac{K}{K-k_1}}z-\sqrt{\frac{k_1}{K-k_1}}c_2\Big)^2\Big\}dz$$

$$\leq C\int_{\beta_0}^{\infty}\exp\Big\{-\frac{1}{2}\Big(\sqrt{\frac{K}{K-k_1}}z-\sqrt{\frac{k_1}{K-k_1}}\gamma_m\Big)^2\Big\}dz$$

$$\to 0$$

Combine (i), (ii) and (iii), we have

$$\lim_{m\to\infty}\sup_{\mathcal{M}_1\in\mathfrak{M}}\sup_{\substack{c_1\in[\beta_0,\gamma_m]\\c_2\in[0,\gamma_m]}}\frac{P(\frac{1}{\sqrt{K}}\sum_{i=1}^{K}Z_i>c_2|\frac{1}{\sqrt{|\mathcal{M}_1|}}\sum_{j\in\mathcal{M}_1}Z_j|>c_1)}{P(\frac{1}{\sqrt{K}}\sum_{i=1}^{K}Z_i>c_2)}=0$$

(2)

It is suffice to show

$$\lim_{m\to\infty}\sup_{\mathcal{M}_1\in\mathfrak{M}}\sup_{c_2\in[0,\bar{\Phi}^{-1}(1/m)]}\frac{\mathsf{P}(\frac{1}{\sqrt{K}}\sum_{i=1}^{K}X_i>c_2,\frac{1}{\sqrt{|\mathcal{M}_1|}}\sum_{j\in\mathcal{M}_1}X_j>\beta_0)}{\mathsf{P}(\sum_{i=1}^{K}Z_i'/\sqrt{K}>c_2)}\leq 0$$

Let $\breve{X}_1=\sum_{i\in\mathcal{M}_1}X_i/\sqrt{k_1}$, $\breve{X}_2=\sum_{i\in\mathfrak{M}\setminus\mathcal{M}_1}X_i/\sqrt{K-k_1}$.

Based on lemma 2.2, $\delta_{6m}=|P(\breve{X}_j>p)/P(Z_j>p)-1|\to 0$ uniformly for $j=1,2$ and $p>\alpha_m$.

Thus, uniformly,

$$
\begin{aligned}
&P(\sqrt{\frac{k_1}{K}}\breve{X}_1+\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2,\breve{X}_1>\beta_0)\\
=&P(\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2-\sqrt{\frac{k_1}{K}}\beta_0,\breve{X}_1>\beta_0)\\
&+P(\sqrt{\frac{K-k_1}{K}}\breve{X}_2<c_2-\sqrt{\frac{k_1}{K}}\beta_0,\sqrt{\frac{k_1}{K}}\breve{X}_1+\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2)\\
\le&(1+\delta_{6m})\big[P(\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2-\sqrt{\frac{k_1}{K}}\beta_0,Z_1>\beta_0)\\
&+P(\sqrt{\frac{K-k_1}{K}}\breve{X}_2<c_2-\sqrt{\frac{k_1}{K}}\beta_0,\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}\breve{X}_2>c_2)+P(Z_1>\bar{\Phi}^{-1}(\alpha_m))\big]\\
\le&(1+\delta_{6m})^2\big[P(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0)\big]+(1+\delta_{6m})\sum_{j=1}^{2}P(Z_j>\bar{\Phi}^{-1}(\alpha_m))\\
\le&(1+\delta_{6m})^2\big[P(\sqrt{\frac{k_1}{K}}Z_1+\sqrt{\frac{K-k_1}{K}}Z_2>c_2,Z_1>\beta_0)\big]+2(1+\delta_{6m})\alpha_m\\
\le&o(\mathsf{P}(\sum_{i=1}^{K}Z_i'/\sqrt{K}>c_2))
\end{aligned}
$$

□

Proof of Lemma 2.4. **(i) Prove that (1) can leads to (2):**

On $\cap_{t=1}^{\ell}\mathcal{X}^{(t)}$,

$$\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|I(T_S<\hat{t}^{(\ell)})\le\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}+\Big\{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}\Big\}\epsilon$$

Combined with

$$\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}\le\alpha\sum_{S\in\mathcal{B}^{(\ell)}}|S|\mathbb{I}\{T_S<\hat{t}^{(\ell)}\}$$

and

$$\sum_{S\in\mathcal{B}^{(\ell)}}|S|\mathbb{I}\{T_S<\hat{t}^{(\ell)}\}=\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\mathbb{I}\{T_S<\hat{t}^{(\ell)}\}+\sum_{S\in\mathcal{B}_1^{(\ell)}}|S|\mathbb{I}\{T_S^{(\ell)}\le\hat{t}^{(\ell)}\}$$

$$\le\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\mathbb{I}\{T_S^{(\ell)}<\hat{t}^{(\ell)}\}+Cm^{r_1}$$

We have:

$$(1-\alpha-\alpha\epsilon)\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}\le\alpha Cm^{r_1}$$

Thus, $2|\mathcal{B}_0^{(\ell)}|\hat{t}^{(\ell)}\le\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|\hat{t}^{(\ell)}\le\frac{\alpha}{1-\alpha-\alpha\epsilon}m^{r_1}$, for any $1\le\ell\le L$.

When $\ell=1$, by $|\mathcal{B}_0^{(1)}|=m_0=m(1+o(1))$, we have $\hat{t}^{(\ell)}\le Cm^{(r_1-1)}$.

When $\ell\ge2$, on $\cap_{k=1}^{(\ell)}\mathcal{X}^{(k)}$, we have

$$\max_{k=1,\dots,\ell}\{FDP^{(k)}-\alpha\}<\epsilon$$

which leads to $|\mathcal{B}_0^{(\ell)}|/|\mathcal{B}^{(\ell)}|\to1$. And accordingly, $\hat{t}^{(\ell)}\le Cm^{(r_1-1)}$.

(ii) Prove that statement (2) leads to statement (3)

On layer 1, $\bar{\Phi}(\hat{c}_S)=\hat{t}^{(1)}\le C(m)^{r_1-1}$. On layer $\ell\ge2$ and $\cap_{h=1}^{\ell}\mathcal{X}^{(h)}$, for all $S\in\mathcal{B}^{(\ell)}$,

$$\bar{\Phi}(\hat{c}_S)\le G_S(\hat{c}_S)+\sum_{S'\in\mathcal{U}(S)}\bar{\Phi}(\hat{c}_{S'})\tag{A.2}$$

Suppose $\bar{\Phi}(\hat{c}_{S'})\le C(m)^{r_1-1}$ for $S'\in\cup_{k=1}^{\ell-1}\mathcal{B}^{(k)}$, then together with $G_S(\hat{c}_S)=\hat{t}^{(\ell)}\le Cm^{r_1-1}$ and (A.2), we have

$$\hat{c}_S\ge\sqrt{2(1-r_1)\log m-2\log\log m}=\beta_m$$

for all $S\in\mathcal{B}^{(\ell)}$.

In addition, for $S\in\mathcal{B}^{(\ell)}$, on $\cap_{h=1}^{\ell-1}\mathcal{X}^{(h)}$,

$$G_S(\hat{c}_S)[1-\bar{\Phi}(\frac{\beta_0}{\sqrt{M^{L-1}}})]^{M^{L-1}}\le\bar{\Phi}(\hat{c}_S)\tag{A.3}$$

So we have $\bar{\Phi}(\hat{c}_S) \geq \hat{t}^{(\ell)}(1+o(1))$, and accordingly, $\hat{c}_S \leq \gamma_m$.

Note that the $\hat{c}_S \leq \gamma_m$ only depends on the statement (2) on layer $\ell - 1$. Thus, we can apply the conclusion to show $\mathsf{P}(m_0\hat{t}^{(\ell)} > c\log m) \to 1$ in the proof of theorem 1.

(iii) Prove that statement (1) holds on layer 1 ($\ell = 1$):

Define $\nu_m = [(|\mathcal{A}'|^2/m + \delta_{2m}) \vee 1]/\sqrt{c_{\mathrm{md}} \log m}$. Let $0 = c_0 < ... < c_{\lceil \gamma_m/\nu_m \rceil} = \gamma_m$ satisfy $c_k - c_{k-1} = \nu_m$ for $1 \leq k < \lceil \gamma_m/\nu_m \rceil$ and $c_{\lceil \gamma_m/\nu_m \rceil} - c_{\lceil \gamma_m/\nu_m \rceil - 1} \leq \nu_m$. We can get the corresponding p-values sequence $q_0 > ... > q_{\lceil \gamma_m/\nu_m \rceil}$ with $q_k = 1 - \Phi(c_k)$. Let value $q^{(1)} = C^{(1)} c_{\mathrm{md}}/m$, by (2.23), we have $P(\hat{t} > q^{(1)}) \to 1$. We define the working p-value sequence on layer 1 as $P^{(1)}_{sub} = \{q_0, ..., q_{k^{(1)}}, q^{(1)}\}$, where $k^{(1)} \in \{0, ..., \lceil \gamma_m/\nu_m \rceil - 1\}$ is the index s.t. $q_{k^{(1)}} \geq q^{(1)}$ and $q_{k^{(1)}+1} \leq q^{(1)}$.

If $\forall \epsilon > 0$,

$$P\left(\max_{q \in P^{(1)}_{sub}} \left| \frac{\sum_{S \in \mathcal{B}^{(1)}_0} I(X_S > \bar{\Phi}^{-1}(q)) - \sum_{S \in \mathcal{B}^{(1)}_0} P(X_S > \bar{\Phi}^{-1}(q))(1 - \delta_{0m})}{\sum_{S \in \mathcal{B}^{(1)}_0} q} \right| > \epsilon \right) \to 0 \tag{A.4}$$

Then,

$$\begin{aligned}
&P\left(\max_{q \in P^{(1)}_{sub}} \frac{\sum_{S \in \mathcal{B}^{(1)}_0} I(T^{(1)}_S < q) - \sum_{S \in \mathcal{B}^{(1)}_0} q}{\sum_{S \in \mathcal{B}^{(1)}_0} q} > \epsilon \right) \\
\leq &P\left(\max_{q \in P^{(1)}_{sub}} \frac{\sum_{S \in \mathcal{B}^{(1)}_0} I(X_S > \bar{\Phi}^{-1}(q)) - \sum_{S \in \mathcal{B}^{(1)}_0} P(\tilde{X}_S > \bar{\Phi}^{-1}(q))}{\sum_{S \in \mathcal{B}^{(1)}_0} q} > \epsilon \right) \\
\leq &P\left(\max_{q \in P^{(1)}_{sub}} \frac{\sum_{S \in \mathcal{B}^{(1)}_0} I(X_S > \bar{\Phi}^{-1}(q)) - \sum_{S \in \mathcal{B}^{(1)}_0} P(X_S > \bar{\Phi}^{-1}(q))(1 - \delta_{0m})}{\sum_{S \in \mathcal{B}^{(1)}_0} q} > \epsilon \right) \\
\leq &P\left(\max_{q \in P^{(1)}_{sub}} \left| \frac{\sum_{S \in \mathcal{B}^{(1)}_0} I(X_S > \bar{\Phi}^{-1}(q)) - \sum_{S \in \mathcal{B}^{(1)}_0} P(X_S > \bar{\Phi}^{-1}(q))(1 - \delta_{0m})}{\sum_{S \in \mathcal{B}^{(1)}_0} q} \right| > \epsilon \right) \\
= &o(1)
\end{aligned} \tag{A.5}$$

Together with the fact that $\sup_{j=1,...,k}\left|q_{(j)}/q_{(j-1)}-1\right|=o(1)$, we have

$$P\bigg(\sup_{q\in[q^{(1)},\alpha]}\frac{\sum_{S\in\mathcal{B}_0^{(1)}}I(T_S<q)-\sum_{S\in\mathcal{B}_0^{(1)}}q}{\sum_{S\in\mathcal{B}_0^{(1)}}q}>\epsilon\bigg)=o(1)$$

Thus, to prove (1) holds on layer 1, we only need to show (A.4).

Define $C_{sub}^{(1)}=\{c_0,...,c_{k'},c'\}$, with $c'=\bar{\Phi}^{-1}(q')$. In order to show (A.4), it is suffice to show

$$\int_0^{c'}P\bigg\{\bigg|\frac{\sum_{S\in\mathcal{B}_0^{(1)}}I(X_S>c)-P(X_S>c)(1-\delta_{0m})}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)}\bigg|\geq\epsilon\bigg\}dc=o(\nu_m)\tag{A.6}$$

Note that by Markov inequality,

$$\begin{aligned}&P\bigg\{\bigg|\frac{\sum_{S\in\mathcal{B}_0^{(1)}}\big[I(X_S>c)-P(X_S>c)(1-\delta_{0m})\big]}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)}\bigg|\geq\epsilon\bigg\}\\ \leq&P\bigg\{\bigg|\frac{\sum_{S\in\mathcal{B}_0^{(1)}}\big[I(X_S>c)-P(X_S>c)\big]}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)}\bigg|\geq\epsilon-(1+\delta_{0m})\delta_{0m}\bigg\}\\ \leq&\frac{\sum_{S,S'\in\mathcal{B}_0^{(1)}}\big[P(X_S>c,X_{S'}>c)-P(X_S>c)P(X_{S'}>c)\big]}{\big(\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)\big)^2[\epsilon-(1+\delta_{0m})\delta_{0m}]^2}\end{aligned}$$

We can divide the $S,S'\in\mathcal{B}_0^{(1)}$ into the following three subsets:

$$\begin{aligned}\mathcal{B}_{01}^{(1)}&=\{S,S'\in\mathcal{B}_0^{(1)}:S=S'\}\\ \mathcal{B}_{02}^{(1)}&=\{S,S'\in\mathcal{B}_0^{(\ell)}:S\neq S',\exists A,A'\in\mathcal{A}^{(L)},s.t.S\subset A,S'\subset A',\text{ and }A'\in\Gamma_A\}\\ \mathcal{B}_{03}^{(1)}&=\{S,S'\in\mathcal{B}_0^{(1)}:S\neq S'\}\setminus\mathcal{B}_{02}^{(1)}\end{aligned}\tag{A.7}$$

Then,

$$\frac{\sum_{(S,S')\in\mathcal{B}_{01}^{(1)}}\big[P(X_S>c,X_{S'}>c)-P(X_S>c)P(X_{S'}>c)\big]}{\big(\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)\big)^2[\epsilon-(1+\delta_{0m})\delta_{0m}]^2}\leq\frac{C}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)}$$

Based on condition 2.3,

$$\frac{\sum_{(S,S')\in\mathcal{B}_{02}^{(1)}}\left[P(X_S>c,X_{S'}>c)-P(X_S>c)P(X_{S'}>c)\right]}{\left(\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)\right)^2[\epsilon-(1+\delta_{0m})\delta_{0m}]^2}\leq\frac{C(|\mathcal{A}'|^2/m+\delta_{2m})}{\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)}$$

In addition,

$$\frac{\sum_{(S,S')\in\mathcal{B}_{03}^{(1)}}\left[P(X_S>c,X_{S'}>c)-P(X_S>c)P(X_{S'}>c)\right]}{\left(\sum_{S\in\mathcal{B}_0^{(1)}}\bar{\Phi}(c)\right)^2[\epsilon-(1+\delta_{0m})\delta_{0m}]^2}=o(1)$$

Thus, after some calculation, we can prove (A.6) and then $\mathsf{P}(\mathcal{X}^{(1)})\to 1$.

Similarly, if $|\tilde{\Omega}_0|=m(1+o(1))$, based on (A.4), we have

$$P\left(\max_{q\in P_{sub}^{(1)}}\left|\frac{\sum_{S\in\mathcal{B}_0^{(1)}}I(T_S^{(1)}<q)-\sum_{S\in\mathcal{B}_0^{(1)}}q}{\sum_{S\in\mathcal{B}_0^{(1)}}q}\right|>\epsilon\right)=o(1)$$

Hence, $\mathsf{P}(\mathcal{X}'^{(1)})\to 1$.

(iv) Prove that statement (1) holds on layer $\ell\geq 2$ when statement (1) holds on previous layers:

On layer ℓ, we can divide the $S,S'\in\mathcal{B}_0^{(\ell)}$ into the following three subsets:

$$\mathcal{B}_{01}^{(\ell)}=\{S,S'\in\mathcal{B}_0^{(\ell)}:S=S',\{T_i:i\in S\}\text{ are mutually independent}\}$$

$$\mathcal{B}_{02}^{(\ell)}=\{S,S'\in\mathcal{B}_0^{(\ell)}:\exists A,A'\in\mathcal{A}^{(L)},s.t.S\subset A,S'\subset A',\text{ and }A'\in\Gamma_A\}$$

$$\mathcal{B}_{03}^{(\ell)}=\{S,S'\in\mathcal{B}_0^{(\ell)}:S\neq S'\}\setminus\mathcal{B}_{02}^{(\ell)}$$

Consider the p-values sequence $q_0>...>q_{\lceil\gamma_m/\nu_m\rceil}$ constructed in (iii). Let $q^{(\ell)}=C^{(\ell)}c_{\mathrm{md}}/m$, by (2.23), we have $P(\hat{t}>q^{(\ell)})\to 1$. We define the working p-value sequence on layer 1 as $P_{sub}^{(\ell)}=\{q_0,...,q_{k^{(\ell)}},q^{(\ell)}\}$, where $k^{(\ell)}\in\{0,...,\lceil\gamma_m/\nu_m\rceil-1\}$ is the index s.t. $q_{k^{(\ell)}}\geq q^{(\ell)}$ and $q_{k^{(\ell)}+1}\leq q^{(\ell)}$.

In view of statement (3) and Lemma 2.3, we have

$$\sup_{k=0,...,\lceil\gamma_m/\nu_m\rceil}\left|\frac{G_S(c_k)}{\bar{\Phi}(c_k)}-1\right|=o(1)$$

Together with statement (3) and Lemma 2.2, there exists $\delta_5(m) \to 0$ with

$$\begin{aligned}
&\max_{S\in\mathcal{B}_0^{(\ell)}} \frac{\mathsf{P}(X_S > \bar{\Phi}^{-1}(q)|\mathcal{Q}^{(1:\ell-1)})}{q} \\
\leq &\max_{S\in\mathcal{B}_0^{(\ell)}} \frac{P(X_S > \bar{\Phi}^{-1}(q))}{P(Z_S > \bar{\Phi}^{-1}(q))[1-\bar{\Phi}(\frac{\beta_0}{\sqrt{M^{L-1}}})]^{M^{L-1}}} \\
\leq &1+\delta_5(m)
\end{aligned}$$

Then $\forall \epsilon > 0$, by following the similar arguments in (iii), we can have

$$P\bigg(\max_{q\in P_{sub}^{(\ell)}} \bigg|\frac{\sum_{S\in\mathcal{B}_{01}^{(\ell)}}|S|I(X_S > \bar{\Phi}^{-1}(q)) - \sum_{S\in\mathcal{B}_{01}^{(\ell)}}|S|P(X_S > \bar{\Phi}^{-1}(q)|\mathcal{Q}^{(1:\ell-1)})(1+\delta_{0m})}{\sum_{S\in\mathcal{B}_{01}^{(\ell)}}|S|q}\bigg| > \epsilon \bigg| \mathcal{Q}^{(1:\ell-1)}$$

$$\to 0 \tag{A.8}$$

Then,

$$\begin{aligned}
&P\bigg(\max_{q\in P_{sub}^{(\ell)}} \frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|I(T_S < q) - \sum_{S\in\mathcal{B}_0^{(\ell)}}|S|q}{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|q} > \epsilon \bigg| \mathcal{Q}^{(1:\ell-1)}\bigg) \\
\leq &P\bigg(\max_{q\in P_{sub}^{(\ell)}} \frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|I(X_S > \bar{\Phi}^{-1}(q)) - \sum_{S\in\mathcal{B}_0^{(\ell)}}|S|P(X_S > \bar{\Phi}^{-1}(q)|\mathcal{Q}^{(1:\ell-1)})}{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|q} > \epsilon/2 \bigg| \mathcal{Q}^{(1:\ell-1)}\bigg) \\
= &o(1)
\end{aligned} \tag{A.9}$$

Together with the fact that $\sup_{j=1,\ldots,k}\left|q_{(j)}/q_{(j-1)} - 1\right| = o(1)$, we have

$$P\bigg(\sup_{q\in[q^{(\ell)},\alpha]} \frac{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|I(T_S < q) - \sum_{S\in\mathcal{B}_0^{(\ell)}}|S|q}{\sum_{S\in\mathcal{B}_0^{(\ell)}}|S|q} > \epsilon \bigg| \mathcal{Q}^{(1:\ell-1)}\bigg) = o(1)$$

And thus $P(\mathcal{X}^{(\ell)}) \to 1$.

Similarly, based on Lemma 2.3 (2) , when $|\tilde{\Omega}_0| = m(1+o(1))$, we have $P(\mathcal{X}'^{(\ell)}) \to 1$. □

Proof of Lemma 2.5. When $\ell = 1$:

for $\delta = 1/m^4$,

$$\begin{aligned}\sum_{S\in\mathcal{B}_0^{(1)}} |S|\hat{t}^{(1)} &\le \alpha \sum_{S\in\mathcal{B}^{(1)}} |S| I(T_S < \hat{t}^{(1)}) \\ &\le \alpha \sum_{S\in\mathcal{B}^{(1)}} |S| I(T_S < \hat{t}^{(1)} + \delta) \\ &\le \sum_{S\in\mathcal{B}_0^{(1)}} |S|\hat{t}^{(1)}(1+o(1)) \qquad \text{(A.10)}\end{aligned}$$

Assume (2.19) holds on layer $1, \ldots, \ell - 1$. Then,

$$\sum_{S\in\mathcal{B}_0^{(\ell)}} |S|\hat{t}^{(\ell)} \le \alpha(1+o(1)) \sum_{S\in\mathcal{B}^{(\ell)}} |S| I(T_S < \hat{t}^{(\ell)})$$

Thus, by following the similar arguments on (A.10), we can get (2.19) on layer ℓ.

□

Appendix B

Appendix for Chapter 3

B.1 Proof of Lemmas

Proof of Lemma 3.1. Consider a partition on the sample space Ω with m bins $\{\Omega_1, \ldots, \Omega_m\}$. On such a partition, N_1 samples from cohort 1 and N_2 samples from cohort 2 are collected. The total number of samples falling into bin i is n_i, $i = 1, \ldots, m$, with $\tilde{X}_i$ from cohort 1 and X_i from cohort 2. Without loss of generality, let's consider the joint distribution of $(X_1, \ldots, X_K)$ given $n_1, \ldots, n_K$ and N_1, N_2.

By Bayes's equality,

$$
\begin{aligned}
&\mathsf{P}(X_1 = x'_1, \ldots, X_K = x'_K \mid n_1, \ldots, n_K, N_1, N_2) \\
=&\frac{\mathsf{P}(X_1 = x'_1, \ldots, X_K = x'_K, n_1, \ldots, n_K \mid N_1, N_2)}{\mathsf{P}(n_1, \ldots, n_K \mid N_1, N_2)} \\
=&\frac{\mathsf{P}(\tilde{X}_1 = x'_1, \ldots, \tilde{X}_K = \tilde{x}'_K, X_1 = x'_1, \ldots, X_K = x'_K, \mid N_1, N_2)}{\sum_{\substack{x_1+\tilde{x}_1=n_1 \\ \vdots \\ x_K+\tilde{x}_K=n_K}} \mathsf{P}(X_1 = x_1, \ldots, X_K = x_K, \tilde{X}_1 = x_1, \ldots, \tilde{X}_K = \tilde{x}_K \mid N_1, N_2)}
\end{aligned} \tag{B.1}
$$

It is equivalent to consider that N_1/N_2 samples have been partitioned into $K + 1$

bins with bin i containing $\tilde{X}_i/X_i$ samples such that $\tilde{X}_i + X_i = n_i$, and $\sum_{i=1}^{K+1} \tilde{X}_i = N_1$ and $\sum_{i=1}^{K} X_i = N_2$. For bin i with $i \in \{1, \ldots, K\}$, let $\tilde{p}_i = \int_{\Omega_i} f_1(y)\,\mathrm{d}y$ and $p_i = \int_{\Omega_i} f_2(y)\,\mathrm{d}y$, where f_1 and f_2 are PDF of the two cohorts. Also, let $\tilde{p}_{K+1} = 1 - \sum_{i=1}^{K} \int_{\Omega_i} f_1(y)\,\mathrm{d}y$, and $p_{K+1} = 1 - \sum_{i=1}^{K} \int_{\Omega_i} f_2(y)\,\mathrm{d}y$. Further, let

$$q_i = \frac{N_1}{N}\tilde{p}_i + \frac{N_2}{N}p_i, \quad \tilde{\theta}_i = \frac{\frac{N_1}{N}\tilde{p}_i}{q_i}, \quad \theta_i = \frac{\frac{N_2}{N}p_i}{q_i}$$

It is easy to see that when $M_1 \leq inf_{s\in\{1,2\}} \inf_y f_s(y) \leq \sup_{s\in\{1,2\}} \sup_y f_s(y) \leq M_2$, $\sum_{i=1}^{K} \leq CKn^{(1)}/N$ and $\sum_{i=1}^{K} \tilde{p}_i \leq CKn^{(1)}/N$.

The definition of $\theta_1, \ldots, \theta_K$ are the same as in (3.1). Also we have $\theta_i + \tilde{\theta}_i = 1$, for $i = 1, \ldots, K+1$.

$$\begin{aligned}
&\mathsf{P}(X_1 = x_1, \ldots, X_K = x_K, \tilde{X}_1 = x_1, \ldots, \tilde{X}_K = x_K \mid N_1, N_2) \\
=&\frac{N_1!}{\prod_{i=1}^{K+1}\tilde{x}_i!}\prod_{i=1}^{K+1}\tilde{p}_i^{\tilde{x}_i} \cdot \frac{N_2!}{\prod_{i=1}^{K+1}x_i!}\prod_{i=1}^{K+1}p_i^{x_i} \\
=&\frac{N_1!}{\prod_{i=1}^{K+1}\tilde{x}_i!}\prod_{i=1}^{K+1}\left(\frac{N_1}{N}\tilde{p}_i\right)^{\tilde{x}_i} \cdot \frac{N_2!}{\prod_{i=1}^{K+1}x_i!}\prod_{i=1}^{K+1}\left(\frac{N_2}{N}p_i\right)^{x_i} \cdot \left(\frac{N}{N_1}\right)^{N_1}\left(\frac{N}{N_2}\right)^{N_2} \\
=&\frac{N_1!}{\prod_{i=1}^{K+1}\tilde{x}_i!}\prod_{i=1}^{K+1}\tilde{\theta}_i^{\tilde{x}_i} \cdot \frac{N_2!}{\prod_{i=1}^{K+1}x_i!}\prod_{i=1}^{K+1}\theta_i^{x_i} \cdot \left(\frac{N}{N_1}\right)^{N_1}\left(\frac{N}{N_2}\right)^{N_2} \cdot \prod_{i=1}^{K+1}q_i^{n_i} \\
=&\left(\frac{N}{N_1}\right)^{N_1}\left(\frac{N}{N_2}\right)^{N_2} N_1!N_2! \prod_{i=1}^{K+1}q_i^{n_i} \cdot \frac{1}{\prod_{i=1}^{K}\tilde{x}_i!}\prod_{i=1}^{K}\tilde{\theta}_i^{\tilde{x}_i} \cdot \frac{1}{\prod_{i=1}^{K}x_i!}\prod_{i=1}^{K}\theta_i^{x_i} \cdot \frac{\tilde{\theta}_{K+1}^{\tilde{x}_{K+1}}\theta_{K+1}^{x_{K+1}}}{\tilde{x}_{K+1}!x_{K+1}!}
\end{aligned}$$

Now let

$$g(\tilde{x}_1, \ldots, \tilde{x}_K, x_1, \ldots, x_K) = \frac{1}{\prod_{i=1}^{K}\tilde{x}_i!}\prod_{i=1}^{K}\tilde{\theta}_i^{\tilde{x}_i} \cdot \frac{1}{\prod_{i=1}^{K}x_i!}\prod_{i=1}^{K}\theta_i^{x_i},$$

$$w\{(\tilde{x}_i, x_i, \tilde{x}_i', x_i')_{i=1\ldots,m}\} = \frac{\tilde{x}_{K+1}'!x_{K+1}'!}{\tilde{x}_{K+1}!x_{K+1}!}\frac{\tilde{\theta}_{K+1}^{\tilde{x}_{K+1}}\theta_{K+1}^{x_{K+1}}}{\tilde{\theta}_{K+1}^{\tilde{x}_{K+1}'}\theta_{K+1}^{x_{K+1}'}}.$$

Then,

$$(\text{B.1}) = \frac{g(\tilde{x}'_1, \ldots, \tilde{x}'_K, x'_1, \ldots, x'_K)}{\sum_{\substack{x_1+\tilde{x}_1=n_1 \\ \vdots \\ x_K+\tilde{x}_K=n_K}} g(\tilde{x}_1, \ldots, \tilde{x}_K, x_1, \ldots, x_K) w(\tilde{x}_1, \ldots, \tilde{x}_K, x_1, \ldots, x_K)}$$

We now show that the term $w\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\}$ is close to 1. Let

$$\log w\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\} = A\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\} + B\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\},$$

where

$$A\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\} = \log\left(\frac{\tilde{x}'_{K+1}!}{\tilde{x}_{K+1}!}\right) + \log\left(\frac{x'_{K+1}!}{x_{K+1}!}\right)$$

$$B\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\} = (\tilde{x}_{K+1} - \tilde{x}'_{K+1}) \log \tilde{\theta}_{K+1} + (x_{K+1} - x'_{K+1}) \log \theta_{K+1}$$

Let $\delta_{K+1} = \tilde{x}_{K+1} - \tilde{x}'_{K+1}$. Because $n_{K+1} = \tilde{x}'_{K+1} + x'_{K+1} = \tilde{x}_{K+1} + x_{K+1}$, we have $x'_{K+1} - x_{K+1} = \delta_{K+1}$. It is also easy to see that $|\delta_{K+1}| \leq Kn^{(1)}$.

When $\delta_{K+1} = 0$, $w\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\} = 1$.

Now assume $\delta_{K+1} > 0$. Then

$$\begin{aligned}
&A\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\} \\
=&-\sum_{i=1}^{\delta_{K+1}} \log(\tilde{x}'_{K+1} + i) + \sum_{i=1}^{\delta_{K+1}} \log(x_{K+1} + i) \\
=&\sum_{i=1}^{\delta_{K+1}} \log\left(\frac{x_{K+1} + i}{\tilde{x}'_{K+1} + i}\right) = \sum_{i=1}^{\delta_{K+1}} \log\left(\frac{N_2 - \sum_{i=1}^K x_i + i}{N_1 - \sum_{i=1}^K \tilde{x}'_i + i}\right) \\
=&\delta_{K+1} \log\left(\frac{N_2}{N_1}\right) + \sum_{i=1}^{\delta_{K+1}} \left\{\log\left(1 - \frac{\sum_{i=1}^K x_i + i}{N_2}\right) + \log\left(1 + \frac{\sum_{i=1}^K \tilde{x}'_i - i}{N_1 - \sum_{i=1}^K \tilde{x}'_i + i}\right)\right\}.
\end{aligned}$$

By Taylor expansions, we know that $|\log(1+x) - x| \leq Cx^2$ and $|\log(1-x) + x| \leq Cx^2$. Plugging the expansion bound into the above expression, we have

$$A\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\} = \delta_{K+1} \log\left(\frac{N_2}{N_1}\right) \pm CK(n^{(1)})^2/N. \tag{B.2}$$

When $\delta_{K+1} < 0$, We can show (B.2) similarly.

On the other hand,

$$\begin{aligned} B\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\} =& \delta_{K+1} \log\left(\frac{\tilde{\theta}_{K+1}}{\theta_{K+1}}\right) \\ =& \delta_{K+1} \log\left(\frac{N_1}{N_2}\right) + \delta_{K+1} \log\left(\frac{1-\sum_{i=1}^K \tilde{p}_i}{1-\sum_{i=1}^K p_i}\right) \\ =& \delta_{K+1} \log\left(\frac{N_1}{N_2}\right) \pm CK^2(n^{(1)})^2/N. \end{aligned}$$

Combining the bounding results for $A\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\}$ and $B\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\}$, we have

$$|w\{(\tilde{x}_i, x_i, \tilde{x}'_i, x'_i)_{i=1\ldots,m}\}| \simeq \exp\{C(n^{(1)})^2/N\} \simeq 1 + CK^2(n^{(1)})^2/N.$$

This leads to

$$\text{(B.1)} \simeq \frac{g(\tilde{x}'_1, \ldots, \tilde{x}'_K, x'_1, \ldots, x'_K)}{\sum_{\substack{x_1+\tilde{x}_1=n_1 \\ \vdots \\ x_K+\tilde{x}_K=n_K}} g(\tilde{x}_1, \ldots, \tilde{x}_K, x_1, \ldots, x_K)} \cdot \left\{1 + CK^2(n^{(1)})^2/N\right\}.$$

Then,

$$\sum_{\substack{x_1+\tilde{x}_1=n_1 \\ \vdots \\ x_K+\tilde{x}_K=n_K}} g(\tilde{x}_1, \ldots, \tilde{x}_K, x_1, \ldots, x_K) \simeq \prod_{i=1}^K \sum_{\tilde{x}_i+x_i=n_i} \frac{\tilde{\theta}_i^{\tilde{x}_i}\theta_i^{x_i}}{x_i!\tilde{x}_i!} = \prod_{i=1}^K \frac{1}{n_i!}.$$

Therefore

$$\text{(B.1)} \simeq \prod_{i=1}^K \frac{n_i!}{x_i!\tilde{x}_i!}\tilde{\theta}_i^{\tilde{x}_i}\theta_i^{x_i} \cdot \left\{1 + C(n^{(1)})^2/N\right\}.$$

□

Proof of Lemma 3.2. Note that $\theta_i = \frac{N_2 \int_{\Omega_i} f_2(y)\,\mathrm{d}y}{N_2 \int_{\Omega_i} f_2(y)\,\mathrm{d}y + N_1 \int_{\Omega_i} f_1(y)\,\mathrm{d}y}$. To prove (3.11), it suffices to show that

$$\left| \frac{\int_{\Omega_j} f_2(y)\,\mathrm{d}y}{\int_{\Omega_j} f_1(y)\,\mathrm{d}y} - \frac{\int_{\Omega_{j-1}} f_2(y)\,\mathrm{d}y}{\int_{\Omega_{j-1}} f_1(y)\,\mathrm{d}y} \right| \leq C/m.$$

Let $y_{s,i,\max} = \arg\max_{\Omega_i} f_s(y)$ and $y_{s,i,\min} = \arg\min_{\Omega_i} f_s(y)$, for $s \in \{1, 2\}$. Then

$$\begin{aligned}
&\left| \frac{\int_{\Omega_i} f_2(y)\,\mathrm{d}y}{\int_{\Omega_i} f_1(y)\,\mathrm{d}y} - \frac{\int_{\Omega_{i-1}} f_2(y)\,\mathrm{d}y}{\int_{\Omega_{i-1}} f_1(y)\,\mathrm{d}y} \right| \\
\leq\ & \left| \frac{f_2(y_{2,i,\max})}{f_1(y_{1,i,\min})} - \frac{f_2(y_{2,i-1,\min})}{f_1(y_{1,i-1,\max})} \right| \vee \left| \frac{f_2(y_{2,i,\min})}{f_1(y_{1,i,\max})} - \frac{f_2(y_{2,i-1,\max})}{f_1(y_{1,i-1,\min})} \right| \\
\leq\ & \frac{M_2}{M_1^2} \max\{|f_2(y_{2,i,\max}) - f_2(y_{2,i-1,\min})|, |f_1(y_{1,i-1,\max}) - f_1(y_{1,i,\min})|, \\
& |f_2(y_{2,i,\min}) - f_2(y_{2,i-1,\max})|, |f_1(y_{1,i-1,\min}) - f_1(y_{1,i,\max})|\}. \qquad \text{(B.3)}
\end{aligned}$$

Based on the multivariate Taylor expansion,

$$\begin{aligned}
|f_2(y_{2,i,\max}) - f_2(y_{2,i-1,\min})| &\leq C \sup_{y \in \Omega} \|\nabla f_2(y)\|_2 \, \|y_{2,i,\max} - y_{2,i-1,\min}\|_2 \\
&\leq C M_3 (\mathsf{Span}(\Omega_{i-1}) + \mathsf{Span}(\Omega_i))^p \leq C/m.
\end{aligned}$$

Similarly, we can show that

$$\begin{aligned}
|f_1(y_{1,i-1,\max}) - f_1(y_{1,i,\min})| &\leq C/m \\
|f_2(y_{2,i,\min}) - f_2(y_{2,i-1,\max})| &\leq C/m \\
|f_1(y_{1,i-1,\min}) - f_1(y_{1,i,\max})| &\leq C/m
\end{aligned}$$

Combined with (B.3), we get the conclusion (3.11). □

Proof of Lemma 3.3. Based on the partition process, for any bin Ω_i,

$$\sum_{k=1}^{N} I(\text{Cell } k \text{ falls into } \Omega_i) = n.$$

Taking expectation on both sides, we have

$$N \cdot \mathsf{P}(\text{Cell } k \text{ falls into } \Omega_i) = \int_{\Omega_i} \{N_2 f_2(y) + N_1 f_1(y)\} \, \mathrm{d}y = n.$$

Combined with the definition of θ_i in (3.1), we have

$$n^{(1)} \sum_{i=1}^{m^{(1)}} \theta_i = N_2 \sum_{i=1}^{n^{(1)}} \int_{\Omega_i} f_2(y) \, \mathrm{d}y = N_2 = N\theta_0.$$

It leads to

$$\sum_{i=1}^{m^{(1)}} \theta_i = \sum_{i \in \mathcal{H}_{\text{nul}}^{(1)}} \theta_i + \sum_{i \in \mathcal{H}_{\text{alt}}^{(1)}} \theta_i = m^{(1)}\theta_0.$$

Now let $\delta = \sup_{i \in \mathcal{H}_{\text{alt}}^{(1)}} (\theta_i - \theta_0)$. Then

$$m_0^{(1)}\theta_0 - \sum_{i \in \mathcal{H}_{\text{nul}}^{(1)}} \theta_i = \sum_{i \in \mathcal{H}_{\text{alt}}^{(1)}} (\theta_i - \theta_0) \leq m_1^{(1)}\delta.$$

Because $\theta_i \leq \theta_0$ in $\mathcal{H}_{\text{nul}}^{(1)}$ and $\mathcal{E}_{\text{nul}}^{(1)} = \mathcal{H}_{\text{nul}}^{(1)} \setminus \mathcal{D}_{\text{nul}}^{(1)} \subseteq \mathcal{H}_{\text{nul}}^{(1)}$,

$$\sum_{i \in \mathcal{E}_{\text{nul}}^{(1)}} (\theta_i - \theta_0) \leq m_1^{(1)}\delta.$$

On $\mathcal{E}_{\text{nul}}^{(1)}$, $\theta_i - \theta_0 \geq (n^{(1)} \log m^{(1)})^{-1}$. Combined with Condition 3.1,

$$\mathsf{Card}(\mathcal{E}_{\text{nul}}^{(1)}) \leq m_1^{(1)}\delta(n^{(1)} \log m^{(1)}) \leq r_2 \delta \{m^{(1)}\}^{\frac{r_4}{1-r_4}+r_1} \log m^{(1)} = o(m^{(1)}).$$

Thus,

$$\frac{\mathsf{Card}(\mathcal{D}_{\text{nul}}^{(1)})}{m_0^{(1)}} = 1 - \frac{\mathsf{Card}(\mathcal{E}_{\text{nul}}^{(1)})}{m_0^{(1)}} = 1 - o(1).$$

□

Proof of Lemma 3.4 b). Based on part a), we have

$$\mathsf{P}(\breve{X}' = k) = \frac{1}{\sqrt{2\pi n'\theta(1-\theta)}} \exp\left(-\frac{(k-n'\theta)^2}{2n'\theta(1-\theta)}\right) \cdot (1+\epsilon_{n'}(k))$$

Thus, there exists $\epsilon'_{n,1} = o(1)$, s.t.

$$\begin{aligned}
&\mathsf{P}(\breve{X}' = k)\sqrt{2\pi n\theta(1-\theta)} \exp\left(\frac{(k-n\theta)^2}{2n\theta(1-\theta)}\right) \\
\leq &\sqrt{\frac{n}{n'}} \exp\left\{\delta_0(m,n)\frac{(k-n\theta)^2}{2n\theta(1-\theta)} + n\delta_0^2(m,n) + \delta_0(m,n)\frac{|k-n\theta|}{\theta(1-\theta)}\right\} \\
\leq &\sqrt{\delta_0(m,n)+1} \exp\left\{\delta_0(m,n)\frac{C^2\log m}{2\theta(1-\theta)} + n\delta_0^2(m,n) + \delta_0(m,n)\frac{C\sqrt{n\log m}}{\theta(1-\theta)}\right\} \\
\leq &1+\epsilon'_{n,1}
\end{aligned}$$

Similarly, by some calculation, there exists $\epsilon'_{n,2} = o(1)$ with

$$\mathsf{P}(\breve{X}' = k)\sqrt{2\pi n\theta(1-\theta)} \exp\left(\frac{(k-n\theta)^2}{2n\theta(1-\theta)}\right) \geq 1-\epsilon'_{n,2}$$

□

Proof of Lemma 3.6. Let $\breve{X}_{S_i^{(\ell)}} = \{\breve{X}_j : j \in S_i^{(\ell)}\}$, with $\breve{X}_j \sim \text{Binom}(n_j, \theta_j)$

For any sample space Ω, we have

$$\begin{aligned}
&\mathsf{P}(\breve{X}_{S_i^{(\ell)}} \in \Omega \mid S_i^{(\ell)} \subseteq \mathcal{D}_{\text{nul}}) - \mathsf{P}(\breve{X}_{S_i^{(\ell)}} \in \Omega \mid \theta_j = \theta_0,\ \forall\ j \in S_i^{(\ell)}) \\
= &\sum_{\breve{x}_{S_i^{(\ell)}} \in \Omega} \binom{n_j}{\breve{x}_j} \theta_j^{\breve{x}_j}(1-\theta_j)^{n_j-\breve{x}_j} - \sum_{\breve{x}_{S_i^{(\ell)}} \in \Omega} \binom{n_j}{\breve{x}_j} \theta_0^{\breve{x}_j}(1-\theta_0)^{n_j-\breve{x}_j} \\
= &\sum_{\breve{x}_{S_i^{(\ell)}} \in \Omega} \binom{n_j}{\breve{x}_j} \theta_0^{\breve{x}_j}(1-\theta_0)^{n_j-\breve{x}_j} A_j, \qquad \text{(B.4)}
\end{aligned}$$

where $A_j = \left(\frac{\theta_j}{\theta_0}\right)^{\breve{x}_j} \left(\frac{1-\theta_j}{1-\theta_0}\right)^{n_j-\breve{x}_j} - 1$.

Let $\theta_j = \theta_0 - a_j$. If $j \in \mathcal{D}_{\text{nul}}^{(1)}$, $0 \le a_j < (n^{(1)} \log m^{(1)})^{-1}$. Thus

$$\begin{aligned}\max_{j \in \mathcal{D}_{\text{nul}}^{(1)}} |A_j| \le & \max_{j \in \mathcal{D}_{\text{nul}}^{(1)}} \left| \left(1 - \frac{a_j}{\theta_0}\right)^{\breve{x}_j} \left(1 + \frac{a_j}{1-\theta_0}\right)^{n_j - \breve{x}_j} - 1 \right| \\ \le & \max_{j \in \mathcal{D}_{\text{nul}}^{(1)}} \left| \left(1 + \frac{a_j}{1-\theta_0}\right)^{n_j} - 1 \right| \vee \left| \left(1 - \frac{a_j}{\theta_0}\right)^{n_j} - 1 \right| \\ = & C(\log m^{(1)})^{-1}.\end{aligned}$$

Combining with (B.4), we have

$$\max_{S_i^{(\ell)} \subseteq \mathcal{D}_{\text{nul}}} \frac{\left| \mathsf{P}(\breve{X}_{S_i^{(\ell)}} \in \Omega_i^{(\ell)} \mid S_i^{(\ell)} \subseteq \mathcal{D}_{\text{nul}}) - \mathsf{P}(\breve{X}_{S_i^{(\ell)}} \in \Omega \mid \theta_j = \theta_0,\ \forall\ j \in S_i^{(\ell)}) \right|}{\mathsf{P}(\breve{X}_{S_i^{(\ell)}} \in \Omega \mid \theta_j = \theta_0,\ \forall\ j \in S_i^{(\ell)})} \le C(\log m^{(1)})^{-1}. \tag{B.5}$$

Let

$$\begin{aligned}\widetilde{PA}_i &= \mathsf{P}(\breve{X}_i^{(\ell)} > b_i^{(\ell)}, \breve{X}_{i1}^{(\ell-1)} \le b_{i1}^{(\ell-1)}, \breve{X}_{i2}^{(\ell-1)} \le b_{i2}^{(\ell-1)} \mid S_i^{(\ell)} \subseteq \mathcal{D}_{\text{nul}}^{(1)}) \\ PA_i &= \mathsf{P}(\breve{X}_i^{(\ell)} > b_i^{(\ell)}, \breve{X}_{i1}^{(\ell-1)} \le b_{i1}^{(\ell-1)}, \breve{X}_{i2}^{(\ell-1)} \le b_{i2}^{(\ell-1)} \mid \theta_j = \theta_0, \forall\ j \in S_i^{(\ell)}) \\ \widetilde{PB}_i &= \mathsf{P}(\breve{X}_{i1}^{(\ell-1)} \le b_{i1}^{(\ell-1)}, \breve{X}_{i2}^{(\ell-1)} \le b_{i2}^{(\ell-1)} \mid S_i^{(\ell)} \subseteq \mathcal{D}_{\text{nul}}^{(1)}), \\ PB_i &= \mathsf{P}(\breve{X}_{i1}^{(\ell-1)} \le b_{i1}^{(\ell-1)}, \breve{X}_{i2}^{(\ell-1)} \le b_{i2}^{(\ell-1)} \mid \theta_j = \theta_0, \forall j \in S_i^{(\ell)}).\end{aligned}$$

By (B.5), we have

$$\widetilde{PA}_i = PA_i(1 + a_i), \quad \widetilde{PB}_i = PB_i(1 + b_i),$$

with

$$\max_{S_i^{(\ell)} \subseteq \mathcal{D}_{\text{nul}}^{(1)}} \{\max(|a_i|, |b_i|)\} \le C(\log m^{(1)})^{-1}. \tag{B.6}$$

Thus,

$$\begin{aligned}\left| \widetilde{G}_i^{(\ell)}(b_i^{(\ell)}; c^{(\ell-1)}) - G_{0,i}^{(\ell)}(b_i^{(\ell)}; c^{(\ell-1)}) \right| &= \left| \frac{\widetilde{PA}_i}{\widetilde{PB}_i} - \frac{PA_i}{PB_i} \right| \\ = \left| \frac{PA_i(1+a_i)}{PB_i(1+b_i)} - \frac{PA_i}{PB_i} \right| &= G_{0,i}^{(\ell)}(b_i^{(\ell)}; c^{(\ell-1)}) \left| \frac{a_i - b_i}{1 + b_i} \right|.\end{aligned}$$

Together with (B.6), we have

$$\max_{S_i^{(\ell)} \subseteq \mathcal{D}_{\text{nul}}} \frac{\left| \widetilde{G}_i^{(\ell)}(b_i^{(\ell)}; c^{(\ell-1)}) - G_{0,i}^{(\ell)}(b_i^{(\ell)}; c^{(\ell-1)}) \right|}{G_{0,i}^{(\ell)}(b_i^{(\ell)}; c^{(\ell-1)})} \leq C(\log m^{(1)})^{-1}$$

□

Proof of Lemma 3.7. Define a set $\mathcal{D}_b = \{b : |b - n_i\theta_0| < 2\sqrt{n_i \log m \theta_0(1-\theta_0)}\}$.

Let $Z \sim N(0,1)$. Based on Lemma 3.4 and Lemma 3.5, we have,

$$\begin{aligned}
&\sup_{b \in \mathcal{D}_b} \left| \frac{G_{0,i}^{(1)}(b; \hat{c}^{(0)})}{P(Z > \frac{b - n_i\theta_0}{\sqrt{n_i\theta_0(1-\theta_0)}})} - 1 \right| \\
\leq &\sup_{b \in \mathcal{D}_b} \left| \frac{P(\frac{\breve{X} - n_i\theta_0}{\sqrt{n_i\theta_0(1-\theta_0)}} > 2\sqrt{\log m}) - P(Z > 2\sqrt{\log m})}{P(Z > \frac{b - n_i\theta_0}{\sqrt{n_i\theta_0(1-\theta_0)}})} \right| \\
&+ \sup_{b \in \mathcal{D}_b} \left| \frac{P(\frac{b - n_i\theta_0}{\sqrt{n_i\theta_0(1-\theta_0)}} < \frac{\breve{X} - n_i\theta_0}{\sqrt{n_i\theta_0(1-\theta_0)}} \leq 2\sqrt{\log m}) - P(2\sqrt{\log m} \geq Z > \frac{b - n_i\theta_0}{\sqrt{n_i\theta_0(1-\theta_0)}})}{P(Z > \frac{b - n_i\theta_0}{\sqrt{n_i\theta_0(1-\theta_0)}})} \right| \\
\to &0
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sup_{b: |b - n_i\theta_0| \leq \sqrt{2n_i \log m \theta_0(1-\theta_0)}} \left| \frac{G_{0,i}^{(1)}(b + o(\sqrt{\frac{n}{\log m}}); \hat{c}^{(0)})}{G_{0,i}^{(1)}(b; \hat{c}^{(0)})} - 1 \right| \\
\leq &\sup_{b: |b - n_i\theta_0| \leq \sqrt{2n_i \log m \theta_0(1-\theta_0)}} \left| \frac{P(Z > \frac{b - n_i\theta_0}{\sqrt{n_i\theta_0(1-\theta_0)}} + o(\sqrt{\frac{1}{\log m}}))}{P(Z > \frac{b - n_i\theta_0}{\sqrt{n_i\theta_0(1-\theta_0)}})} - 1 \right| + o(1) \\
= &o(1)
\end{aligned}$$

When $b \leq n_i\theta_0 - \sqrt{2n_i \log m\theta_0(1-\theta_0)}$, we have

$$\sup_{b:b\leq n_i\theta_0-\sqrt{2n_i \log m\theta_0(1-\theta_0)}} \left| \frac{G_{0,i}^{(1)}(b + o(\sqrt{\frac{n}{\log m}}); \hat{c}^{(0)})}{G_{0,i}^{(1)}(b; \hat{c}^{(0)})} - 1 \right|$$

$$\leq \frac{1 - G_{0,i}^{(1)}(n_i\theta_0 - \sqrt{2n_i \log m\theta_0(1-\theta_0)}; \hat{c}^{(0)})}{G_{0,i}^{(1)}(n_i\theta_0 - \sqrt{2n_i \log m\theta_0(1-\theta_0)}; \hat{c}^{(0)})}$$

$$= o(1)$$

Accordingly, we have

$$\sup_{b:b-n_i\theta_0\leq\sqrt{2n_i \log m\theta_0(1-\theta_0)}} \left| \frac{G_{0,i}^{(1)}(b + o(\sqrt{\frac{n}{\log m}}); \hat{c}^{(0)})}{G_{0,i}^{(1)}(b; \hat{c}^{(0)})} - 1 \right| \to 0$$

□

Proof of Lemma 3.8. Given (b) follows immediately after (a), we will focus on the proof of (a).

Let $Z_1, Z_2, ..., Z_{2^{\ell-1}} \overset{iid}{\sim} N(0,1)$, $\tau^{(\ell)} = \frac{b^{(\ell)} - n^{(\ell)}\theta_0}{\sqrt{\{n^{(\ell)}\theta_0(1-\theta_0)\}}}$, $\tilde{X}_{k_1} = \frac{\sum_{i=1}^{2^{k-1}} \breve{X}_i - n^{(k)}\theta_0}{\sqrt{n^{(k)}\theta_0(1-\theta_0)}}$, $\tilde{X}_{k_2} = \frac{\sum_{i=2^{k-1}+1}^{2^{\ell-1}} \breve{X}_i - (n^{(\ell)} - n^{(k)})\theta_0}{\sqrt{(n^{(\ell)} - n^{(k)})\theta_0(1-\theta_0)}}$. When $\tau_i^{(\ell)} > \beta_0/\sqrt{2^{\ell-k}}$,

$$P(\sum_{i=1}^{2^{\ell-1}} \breve{X}_i > b^{(\ell)}, \sum_{i=1}^{2^{k-1}} \breve{X}_i > b_k)$$

$$= P(\frac{1}{\sqrt{2^{\ell-k}}}\tilde{X}_{k_1} + \sqrt{1 - \frac{1}{2^{\ell-k}}}\tilde{X}_{k_2} > \tau^{(\ell)}, \tilde{X}_{k_1} > \beta_0)$$

$$\leq \mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}\tilde{X}_{k_1} + \sqrt{1 - \frac{1}{2^{\ell-k}}}\tilde{X}_{k_2} > \tau^{(\ell)}, \beta_0 < \tilde{X}_{k_1} \leq \sqrt{2^{\ell-k}}\gamma, -\gamma < \tilde{X}_{k_2} \leq \left[1 - \frac{1}{2^{\ell-k}}\right]^{-1/2}\gamma)$$

$$+ \mathsf{P}(\tilde{X}_{k_1} > \sqrt{2^{\ell-k}}\gamma) + \mathsf{P}(\tilde{X}_{k_2} > \left[1 - \frac{1}{2^{\ell-k}}\right]^{-1/2}\gamma) \tag{B.7}$$

By applying the Lemma 3.4 on the first term in (B.7),

$$\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}\tilde{X}_{k_1}+\sqrt{1-\frac{1}{2^{\ell-k}}}\tilde{X}_{k_2}>\tau^{(\ell)},\beta_0<\tilde{X}_{k_1}\leq\sqrt{2^{\ell-k}}\gamma,-\gamma<\tilde{X}_{k_2}\leq\left[1-\frac{1}{2^{\ell-k}}\right]^{-1/2}\gamma)$$

$$\leq 2\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},\beta_0<Z_1\leq\sqrt{2^{\ell-k}}\gamma,-\gamma<Z_2\leq\left[1-\frac{1}{2^{\ell-k}}\right]^{-1/2}\gamma)$$

$$\leq 2\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)$$

Thus, based on the Lemma 3.5, when m and n are large sufficiently, we have

$$\begin{aligned}&\frac{P(\sum_{i=1}^{2^{\ell-1}}\breve{X}_i>b^{(\ell)},\sum_{i=1}^{2^{k-1}}\breve{X}_i>b_k)}{\mathsf{P}(\sum_{i=1}^{2^{\ell-1}}\breve{X}_i>b^{(\ell)})}\\ \leq&\frac{4\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}+o(1)\end{aligned}\tag{B.8}$$

In addition, when $\tau_i^{(\ell)}\leq\beta_0/\sqrt{2^{\ell-k}}$, we have

$$\begin{aligned}&\max_{k=1,\ldots,\ell-1}\frac{P(\sum_{i=1}^{2^{\ell-1}}\breve{X}_i>b^{(\ell)},\sum_{i=1}^{2^{k-1}}\breve{X}_i>b_k)}{\mathsf{P}(\sum_{i=1}^{2^{\ell-1}}\breve{X}_i>b^{(\ell)})}\\ \leq&\max_{k=1,\ldots,\ell-1}\frac{2\mathsf{P}(Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\beta_0/\sqrt{2^{\ell-k}})}\to 0\end{aligned}$$

Thus, the prove for (3.17) is finished when

$$\lim_{m\to\infty}\max_{k=1,\ldots,\ell-1}\sup_{\tau^{(\ell)}\in(\beta_0/\sqrt{2^{\ell-k}},\gamma]}\frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}=0$$

For each m, define $\mathcal{D}_m=\{\tau^{(\ell)}\in(\beta_0/\sqrt{2^{\ell-k}},\gamma):\frac{d}{d\tau^{(\ell)}}\frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}=$

$0\}$, then

$$\sup_{\tau^{(\ell)}\in(\beta_0/\sqrt{2^{\ell-k}},\gamma]} \frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}$$

$$\leq \max\left\{ \sup_{\tau^{(\ell)}=\beta_0/\sqrt{2^{\ell-k}} \text{ or } \gamma} \frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}, \right.$$

$$\left. \sup_{\tau^{(\ell)}\in\mathcal{D}_m} \frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})} \right\}$$

(i). When $\tau^{(\ell)}=\beta_0/\sqrt{2^{\ell-k}}$ or γ, by L'Hopital's rule, we have

$$\lim_{m\to\infty}\max_{k=1,\ldots,\ell-1} \frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}=0$$

(ii). When $\tau^{(\ell)}\in\mathcal{D}_m$, given

$$\begin{aligned}
0=&\frac{d}{d\tau^{(\ell)}}\frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}\\
=&\frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})\frac{d}{d\tau^{(\ell)}}\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})^2}\\
&-\frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)\frac{d}{d\tau^{(\ell)}}\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})^2}
\end{aligned}$$

We have

$$\frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}$$

$$=\frac{\frac{d}{d\tau^{(\ell)}}\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)},Z_1>\beta_0)}{\frac{d}{d\tau^{(\ell)}}\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1+\sqrt{1-\frac{1}{2^{\ell-k}}}Z_2>\tau^{(\ell)})}$$

Therefore,

$$\max_{k=1,\ldots,\ell-1} \sup_{\tau^{(\ell)}\in\mathcal{D}_m} \frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1 + \sqrt{1-\frac{1}{2^{\ell-k}}}Z_2 > \tau^{(\ell)}, Z_1 > \beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1 + \sqrt{1-\frac{1}{2^{\ell-k}}}Z_2 > \tau^{(\ell)})}$$

$$= \max_{k=1,\ldots,\ell-1} \sup_{\tau^{(\ell)}\in\mathcal{D}_m} \frac{\frac{d}{d\tau^{(\ell)}}\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1 + \sqrt{1-\frac{1}{2^{\ell-k}}}Z_2 > \tau^{(\ell)}, Z_1 > \beta_0)}{\frac{d}{d\tau^{(\ell)}}\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1 + \sqrt{1-\frac{1}{2^{\ell-k}}}Z_2 > \tau^{(\ell)})}$$

$$\to 0$$

By combining (i) and (ii),

$$\lim_{m\to\infty} \max_{k=1,\ldots,\ell-1} \sup_{\tau^{(\ell)}\in(\beta_0/\sqrt{2^{\ell-k}},\gamma]} \frac{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1 + \sqrt{1-\frac{1}{2^{\ell-k}}}Z_2 > \tau^{(\ell)}, Z_1 > \beta_0)}{\mathsf{P}(\frac{1}{\sqrt{2^{\ell-k}}}Z_1 + \sqrt{1-\frac{1}{2^{\ell-k}}}Z_2 > \tau^{(\ell)})} = 0$$

□

Proof of Lemma 3.9. If $\hat{c}^{(\ell)} \le a_N^{(\ell)}$, all statements hold immediately. Now we prove the lemma when $\hat{c}^{(\ell)} > a_N^{(\ell)}$.

Step 1: Prove that on Layer 1, (a) holds

Let $n^{(1)}\theta_0 - \frac{\{n\theta_0(1-\theta_0)\}^{1/2}}{\sqrt{\log m \log\log m}} = b_{-1} < b_0 < b_1 < \ldots < b_{\lceil\gamma(\log m\log\log m)^{1/2}\rceil} = n^{(1)}\theta_0 + \{n^{(1)}\theta_0(1-\theta_0)\}^{1/2}\gamma$ satisfy $b_t - b_{t-1} = \frac{\sqrt{n\theta_0(1-\theta_0)}}{\sqrt{\log m\log\log m}}$ for $0 \le t < \lceil\gamma(\log m\log\log m)^{1/2}\rceil$ and $b_{\lceil\gamma(\log m)^{1/2}\rceil} - b_{\lceil\gamma(\log m\log\log m)^{1/2}\rceil-1} \le \frac{\sqrt{n\theta_0(1-\theta_0)}}{\sqrt{\log m\log\log m}}$. We can get the corresponding p-values sequence $q_{-1}^{(1)} > \ldots > q_{\lceil\gamma(\log m\log\log m)^{1/2}\rceil}^{(1)}$, under Binom$(n,\theta_0)$. Let value $q' = \frac{C^{(1)}\log m}{m}$. Based on (3.22), $P(\hat{c}^{(1)} \ge q') \to 1$. We define the working sequence on layer 1 as $Q_{sub}^{(1)} = \{q_{-1}^{(1)}, \ldots, q_t^{(1)}, q'\}$, where $t \in \{0,\ldots,\lceil\gamma(\log m\log\log m)^{1/2}\rceil - 1\}$ is the index s.t. $q_t^{(1)} \ge q'$ and $q_{t+1}^{(1)} \le q'$.

By Markov Inequality,

$$\mathsf{P}\left[\max_{q\in Q^{(1)}_{sub}}\left|\frac{\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}} I(P^{(1)}_{0,i}<q)-\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}\mathsf{P}(P^{(1)}_{0,i}<q)}{\big\{\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}\mathsf{P}(P^{(1)}_{0,i}<q)\big\}^{1/2}}\right|>\Big\{\log m^{(1)}\Big\}^{1.6/4}\right]$$

$$\leq C(\log m)^{3/2}\frac{E\big[(\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}} I(P^{(1)}_{0,i}<q)-\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}\mathsf{P}(P^{(1)}_{0,i}<q))^4\big]}{\big[\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}\mathsf{P}(P^{(1)}_{0,i}<q)\big]^2(\log m)^{1.6}}$$

$$\leq C(\log m)^{3/2}\frac{\big[m_0q(1-q)\big(1-6q(1-q)\big)+m_0^2q^2(1-q)^2\big]\big(1+C\frac{16n^2}{N}\big)}{m_0^2q^2(1+o(1))(\log m)^{1.6}}\ \text{(Lemma 3.1)}$$

$$\leq C(\log m)^{-0.1} \tag{B.9}$$

Thus, by Lemma 3.3 and (3.22),

$$\mathsf{P}\left[\max_{q\in Q^{(1)}_{sub}}\left|\frac{\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}} I(P^{(1)}_{0,i}<q)-\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}\mathsf{P}(P^{(1)}_{0,i}<q)}{\big\{\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}\mathsf{P}(P^{(1)}_{0,i}<q)\big\}}\right|>\frac{1}{C^{(1)}\big\{\log m^{(1)}\big\}^{0.1}}\right]$$

$$\to 0 \tag{B.10}$$

Given

$$\sup_{j=-1,\ldots,t}\left|\frac{q_j}{q_{j+1}}-1\right|\leq\sup_{j=-1,\ldots,t}\left|\frac{G^{(1)}_{0,i}(b_j;\hat{c}^{(0)})}{G^{(1)}_{0,i}(b_{j+1};\hat{c}^{(0)})}-1\right|+o(1)=o(1)$$

Combined with (3.34), (B.9) and (B.10), there exists $\delta_2(m)=o(1)$, s.t.

$$\mathsf{P}\left[\sup_{q\in[q',1/2]}\left|\frac{\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}} I(P^{(1)}_{0,i}<q)-\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}P(P^{(1)}_{0,i}<q)}{\big\{\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}\mathsf{P}(P^{(1)}_{0,i}<q)\big\}}\right|>\delta_2(m)\right]\to 0$$

and accordingly,

$$\mathsf{P}\left[\left|\frac{\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}} I(P^{(1)}_{0,i}<\hat{c}^{(1)})-\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}\mathsf{P}(P^{(1)}_{0,i}<\hat{c}^{(1)})}{\big\{\sum_{i\in\mathcal{B}^{(1)}_{\text{nul}}}\mathsf{P}(P^{(1)}_{0,i}<\hat{c}^{(1)})\big\}}\right|>\delta_2(m)\right]\to 0 \tag{B.11}$$

Step 2: Prove statement (b)

From (3.7), we have

$$m^{(\ell)}\hat{c}^{(\ell)} \leq \alpha \max\left\{\sum_{i=1}^{m^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}), 1\right\} \tag{B.12}$$

We also know that

$$\begin{aligned}\sum_{i=1}^{m^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}) &= \sum_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}) + \sum_{i\in\mathcal{B}_{\text{alt}}^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}) \\ &\leq \sum_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}) + (m^{(\ell)})^{r_1} \end{aligned} \tag{B.13}$$

On $\cap_{h=1}^{\ell}\mathcal{X}^{(h)}$,

$$\begin{aligned}&\sum_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} I(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}) \\ \leq &\sum_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} \mathsf{P}(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)}) + \left\{\sum_{i\in\mathcal{B}_{\text{nul}}^{(\ell)}} \mathsf{P}(P_{0,i}^{(\ell)} < \hat{c}^{(\ell)})\right\}\delta_2(m) \\ \leq &m_0^{(\ell)}\hat{c}^{(\ell)}(1+\delta_2(m)) \end{aligned} \tag{B.14}$$

By combining (B.12), (B.13) and (B.14),

$$(1-\alpha(1+o(1)))m_0^{(\ell)}\hat{c}^{(\ell)} \leq \alpha(m^{(\ell)})^{r_1}$$

It is easy to see that

$$\hat{c}^{(\ell)} \leq C(m^{(\ell)})^{r_1-1}, \tag{B.15}$$

Step 3: Prove statement (c) on $\cap_{h=1}^{\ell-1}\mathcal{X}^{(h)}$

We start to prove (3.24) by induction. When $\ell = 2$, based on (3.23), $G_{0,i}^{(1)}(\hat{b}_i^{(1)}; \hat{c}^{(0)}) \leq C(m^{(1)})^{r_1-1}$. Then by Lemma 3.5, we know that $\hat{\tau}_i{}^{(1)} \geq \beta\{1+o(1)\}$.

Now suppose $\hat{\tau}_i^{(k)} \geq \beta\{1+o(1)\}$ for $k = 1, \ldots, \ell-2$. On layer $\ell-1$,

$$\begin{aligned} & G_{0,i}^{(\ell-1)}(\hat{b}_i^{(\ell-1)}; \hat{c}^{(0)}) \\ & \leq G_{0,i}^{(\ell-1)}(\hat{b}_i^{(\ell-1)}; \hat{c}^{(\ell-2)})\{1 - \widetilde{G}_{0,i}^{(\ell-2)}(\hat{b}_i^{(\ell-2)}; \hat{c}^{(0)})\}^2 + 3G_{0,i}^{(\ell-2)}(\hat{b}_i^{(\ell-2)}; \hat{c}^{(0)}) \end{aligned} \tag{B.16}$$

by the right-handed side of (B.16) and (B.15), $G_{0,i}^{(\ell-1)}(b_i^{(\ell-1)}; \hat{c}^{(0)}) \leq C(m^{(1)})^{r_1-1} + 3C(m^{(1)})^{r_1-1} \leq C(m^{(1)})^{r_1-1}$. Then by Lemma 3.5, $\hat{\tau}_i^{(\ell-1)} \geq \beta\{1+o(1)\}$.

Step 4: Prove that (a) holds on layer ℓ when it holds on layer $1, ..., \ell-1$

Let $b_{-1} < b_0 < ... < b_{\lceil \gamma(\log m \log\log m)^{1/2} \rceil}$ be the values defined in the proof at Step 1. In view of Lemma 3.8 (b),

$$\sup_{t=0,\ldots,\lceil \gamma(\log m \log\log m)^{1/2} \rceil} \left| \frac{G_{0,i}^{(\ell)}(b_t; \hat{c}^{(\ell-1)})}{G_{0,i}^{(\ell)}(b_{t-1}; \hat{c}^{(\ell-1)})} - 1 \right| = o(1)$$

By following the similar arguments in Step 1, we have $P(\mathcal{X}^{(\ell)}) \to 1$. Thus,

$$P(\cap_{h=1}^{\ell} \mathcal{X}^{(h)}) \geq 1 - o(1)$$

□

Proof of Lemma 3.10. Based on (3.7), we know that

$$\begin{aligned} m_0^{(1)} G_0^{(1)}(\hat{b}_i^{(1)} - 1; \hat{c}^{(0)}) &> \alpha \max\Big\{ \sum_{i=1}^{m^{(\ell)}} I(X_i^{(1)} > \hat{b}_i^{(1)} - 1), 1 \Big\} \\ &\geq \alpha \max\Big\{ \sum_{i=1}^{m^{(\ell)}} I(X_i^{(1)} > \hat{b}_i^{(1)}), 1 \Big\} \end{aligned} \tag{B.17}$$

By Lemma 3.4 and Lemma 3.5, $g_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(0)}) \asymp G_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(0)}) \hat{\tau}_i^{(\ell)} n^{-1/2}$. By (3.7), we know that $\tau_i^{(\ell)} n^{-1/2} \leq C(\log m^{(1)})^{1/2} n^{-1/2} = o(1)$. Combined with (3.20), we know

that on $\mathcal{X}$,

$$\begin{aligned}
&G_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)} - 1; \hat{c}^{(\ell-1)}) \\
\leq &G_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)}) + \frac{g_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(0)})}{\{1 - G_{0,i}^{(\ell-1)}(\hat{b}_{i_1}^{(\ell-1)}; \hat{c}^{(0)})\}\{1 - G_{0,i}^{(\ell-1)}(\hat{b}_{i_2}^{(\ell-1)}; \hat{c}^{(0)})\}} \\
\leq &G_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)}) \left\{1 + C(\log m^{(1)})^{1/2} n^{-1/2}\right\} \qquad \text{(B.18)}
\end{aligned}$$

Here,

$$\begin{aligned}
g_{0,i}^{(\ell)}(b^{(\ell)}; c^{(\ell-1)}) = \mathsf{P}(&\breve{X}_i^{(\ell)} = b^{(\ell)} \mid G_{0,i_1}^{(\ell-1)}(\breve{X}_{i_1}^{(\ell-1)}; \hat{c}^{(\ell-2)}) \geq \hat{c}^{(\ell-1)}, G_{0,i_2}^{(\ell-1)}(\breve{X}_{i_2}^{(\ell-1)}; \hat{c}^{(\ell-2)}) \geq \hat{c}^{(\ell-1)}, \\
&\theta_j = \theta_0, \forall\ j \in S_i^{(\ell)})
\end{aligned}$$

Thus on $\mathcal{X}$,

$$\begin{aligned}
m_0^{(\ell)} G_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)}) &\leq \alpha \max\left\{\sum_{i=1}^{m^{(\ell)}} I(X_i^{(\ell)} > \hat{b}_i^{(\ell)}), 1\right\} \\
&\leq m_0^{(\ell)} G_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)}) \left\{1 + C(\log m^{(1)})^{1/2} n^{-1/2}\right\}.
\end{aligned}$$

And therefore,

$$\frac{m_0^{(\ell)} G_{0,i}^{(\ell)}(\hat{b}_i^{(\ell)}; \hat{c}^{(\ell-1)})}{\max\left\{\sum_{i=1}^{m^{(\ell)}} I(X_i^{(\ell)} > \hat{b}_i^{(\ell)}), 1\right\}} = \alpha \left\{1 + C(\log m^{(1)})^{1/2} n^{-1/2}\right\}.$$

□

Appendix C

Appendix for Chapter 4

Suppose there are N_1 cases and N_0 controls in the data. The total sample size is denoted by $N = N_1 + N_0$. For subject i, let the outcome be $Y_k = \mathbb{I}\{\text{subject } k \text{ is a case}\}$, the non-genetic covariate be $\boldsymbol{Z}_k = (Z_{k,1}, \ldots, Z_{k,p})^T$, and the genetic vector $\boldsymbol{G}_{k,i} = (G_{k,j} : j \in \mathcal{V}(\{i\}))^T$ with $G_{k,j}$ the indicator of whether subject k carries the minor allele of RV j. There are D functional domains. Let FD_d be the set of qualified RVs on domain d, $d \in [D]$.

C.1 Algorithm to construct leaves

We apply Algorithm 6 to partition D functional domains into leaves. Suppose the domains and their RVs are ordered based on their genomic locations. We hope that at least M subjects carrying RV minor alleles on each leaf unless this condition cannot be hold for the entire domain. The outputs of Algorithm 6 are the RV sets of all the

leaves. For leaf $\{i\}$, its RV set is $\mathcal{V}(\{i\})$.

Algorithm 6: Algorithm to partition D functional domains into leaves.

Data: Set of functional domains $\{\mathrm{FD}_d : d \in [D]\}$, leaf size M.
Result: $\mathfrak{V} = \{\mathcal{V}(\{i\})\colon\ i \in [m^{(1)}]\}$.
$\mathfrak{V} = \emptyset$;
for $d \in [D]$ **do**
  $V_0 = \emptyset$; $V_1 = \emptyset$;
  for $j \in \mathrm{FD}_d$ **do**
    if $\sum_{k=1}^{N} \mathbb{I}\{\sum_{j' \in V_1 \cup \{j\}} G_{k,j'} > 0\} < M$ **then**
      if $j + 1 \in \mathrm{FD}_d$ **then**
        $V_1 = V_1 \cup \{j\}$;
      else *// reaching the end of the domain*
        $V_1 = V_1 \cup V_0$; $\mathfrak{V} = (\mathfrak{V} \setminus \{V_0\}) \cup \{V_1\}$; $V_0 = V_1 = \emptyset$;
    else
      if $j + 1 \in \mathrm{FD}_d$ **then**
        $V_1 = V_1 \cup \{j\}$; $\mathfrak{V} = \mathfrak{V} \cup \{V_1\}$; $V_0 = V_1$; $V_1 = \emptyset$;
      else *// reaching the end of the domain*
        $V_1 = V_1 \cup \{j\}$; $\mathfrak{V} = \mathfrak{V} \cup \{V_1\}$; $V_0 = V_1 = \emptyset$;

C.2 Approaches to derive leaf P-value

C.2.1 Lancaster's mid-P correction for the Fisher's exact test (FL)

Let C_i denote the number of cases carrying RVs in leaf i, and c_i be its realized value. When $H_{\{i\}}$ is null,

$$C_i \sim \mathrm{HyperGeom}(M_i, N_1, N_0),$$

where M_i is the number of samples carrying RVs in leaf $\{i\}$. The p-value is given by

$$T_i = \sum_{c:P(C_i=c)<P(C_i=c_i)} \mathsf{P}(C_i = c) + \frac{1}{2} \sum_{c:P(C_i=c)=P(C_i=c_i)} \mathsf{P}(C_k = c)$$

C.2.2 Efficient score statistics with saddle point approximation (SS)

Consider the logistic regression model

$$\text{logit}\{\mathsf{P}(Y_k = 1|\boldsymbol{Z}_k, \boldsymbol{G}_{k,i})\} = \gamma_0 + \boldsymbol{Z}_k^T\boldsymbol{\eta}_i + \boldsymbol{G}_{k,i}^T\boldsymbol{\beta}_i \tag{C.1}$$

Here, $\boldsymbol{\eta}_i$ is the coefficient for non-genetic covariates, and $\boldsymbol{\beta}_i$ is the coefficient for the RVs on leaf $\{i\}$. To test the association between the leaf and the disease, we set the leaf hypothesis $H_{\{i\}}$ as

$$H_{\{i\},0} : \boldsymbol{\beta}_i = 0 \quad \text{versus} \quad H_{\{i\},1} : \exists\ j \in \mathcal{V}(\{i\}),\ \beta_{i,j} \neq 0.$$

We use the collapsing score test to test the hypothesis. Let $\boldsymbol{X}_i = (X_{1,i}, \ldots, X_{N,i})^T$ with $X_{k,i} = \mathbb{I}\{\sum_{j\in\mathcal{V}(\{i\})} G_{k,j} > 0\}$ be the minor allele indicator of leaf $\{i\}$. The score statistic is

$$U_i = \boldsymbol{X}_i^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}),$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ is the outcome vector, and $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_N)^T$ is a vector with $\hat{\mu}_k$ is the estimated probablity of Y_k being a case based on the null model (C.1) with null $\boldsymbol{\beta}_i = 0$. The asymptotic distribution of U_i when leaf $\{i\}$ is null is

$$\frac{U_i}{\sqrt{\boldsymbol{X}_i^T\hat{\boldsymbol{\Sigma}}\boldsymbol{X}_i}} \overset{asym}{\sim} N(0,1)$$

where $\hat{\boldsymbol{\Sigma}}$ is a diagonal matrix with $\hat{\boldsymbol{\Sigma}}_{k,k} = \hat{\mu}_k(1-\hat{\mu}_k)$. The calculation of P-value T_i is adjusted by saddlepoint approximation to better approximate the tails of the normal distribution under the null Dey et al. (2017).

C.2.3 Algorithm to aggregate nodes

We apply Algorithm 7 to aggregate the accepting nodes on layer $(\ell-1)$ to the nodes on layer ℓ, for any later layer $\ell \in \{2, \ldots, L\}$. To simplify notation, suppose the

nodes in $\mathcal{B}'$ are ordered based on their genomic locations. We know that the leaves are nested in the domains. Recursively, assume the nodes on layer $\ell - 1$ are nested in the domains, the algorithm guarantees the nodes on layer ℓ are also nested in the domain. Let $\mathcal{B}'_d = \{S : S \in \mathcal{B}', S \subset \mathrm{FD}_d\}$. For any $i \leq |\mathcal{B}'_d|$, let $S_d(i)$ be the i-th ordered node in $\mathcal{B}'_d$. The orders will be used in Algorithm 7 to aggregate the neighboring leaves.

Algorithm 7: Algorithm to aggregate the accepting nodes on layer $\ell - 1$ to the nodes on layer ℓ

Data: $\{\mathcal{B}'_d : d \in [D]\}$.
Result: $\mathcal{B}^{(\ell+1)}$
$\mathcal{B}^{(\ell+1)} = \emptyset$;
for $d \in [D]$ **do**
  $A = 0$; *// A indicates if the proceeding node is in domian d.*
  $i = 1$;
  while $i \leq |\mathcal{B}'_d|$ **do**
    if $i + 1 \leq |\mathcal{B}'_d|$ **then**
      $\mathcal{B}^{(\ell+1)} = \mathcal{B}^{(\ell+1)} \cup \{S_d(i) \cup S_d(i+1)\}$; $i = i + 2$;
      if $i \leq |\mathcal{B}'_d|$ **then** $A = 1$;
    else if $A = 1$ **then**
      $\mathcal{B}^{(\ell+1)} = \mathcal{B}^{(\ell+1)} \setminus \{S_d(i-1) \cup S_d(i-2)\} \cup \{S_d(i-2) \cup S_d(i-1) \cup S_d(i)\}$;
      $i = i + 1$; $A = 0$;

Bibliography

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010), "A method and server for predicting damaging missense mutations," *Nature methods*, 7, 248–249.

Aitchison, J. (1982), "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, 44, 139–160.

Antoniadis, A., Glad, I. K., and Mohammed, H. (2015), "Local comparison of empirical distributions via nonparametric regression," *Journal of Statistical Computation and Simulation*, 85, 2384–2405.

Asimit, J. L., Day-Williams, A. G., Morris, A. P., and Zeggini, E. (2012), "ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data," *Hum Hered*, 73, 84–94.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010), "Statistical analysis strategies for association studies involving rare variants," *Nature Reviews Genetics*, 11, 773–785.

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)*, 57, 289–300.

Bhatia, G., Bansal, V., Harismendy, O., Schork, N. J., Topol, E. J., Frazer, K., and Bafna, V. (2010), "A covering method for detecting genetic associations between rare variants and common phenotypes," *PLoS Comput Biol*, 6, e1000954.

Biddle, D. A. and Morris, S. B. (2011), "Using Lancaster's mid-P correction to the Fisher's exact test for adverse impact analyses," *J Appl Psychol*, 96, 956–65.

Cai, T. T., Sun, W., and Xia, Y. (2020), "LAWS: A Locally Adaptive Weighting and Screening Approach To Spatial Multiple Testing," *Journal of the American Statistical Association*, pp. 1–30.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016), "DADA2: high-resolution sample inference from Illumina amplicon data," *Nature methods*, 13, 581.

Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'connor, E. M., Cusack, S., Harris, H. M., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., et al. (2012), "Gut microbiota composition correlates with diet and health in the elderly," *Nature*, 488, 178–184.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001), *Introduction To Algorithms*, MIT Press.

Cullup, T., Kho, A. L., Dionisi-Vici, C., Brandmeier, B., Smith, F., Urry, Z., Simpson, M. A., Yau, S., Bertini, E., McClelland, V., et al. (2013), "Recessive mutations in EPG5 cause Vici syndrome, a multisystem disorder with defective autophagy," *Nature genetics*, 45, 83–87.

Daniels, H. E. (1954), "Saddlepoint Approximations in Statistics," *The Annals of Mathematical Statistics*, 25, 631–650.

Dey, R., Schmidt, E. M., Abecasis, G. R., and Lee, S. (2017), "A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS," *Am J Hum Genet*, 101, 37–49.

Dmitrienko, A. and Tamhane, A. C. (2013), "General theory of mixture procedures for gatekeeping," *Biom J*, 55, 402–19.

Duong, T. (2013), "Local significant differences from nonparametric two-sample tests," *Journal of Nonparametric Statistics*, 25, 635–645.

Gelfman, S., Dugger, S., Moreno, C. d. A. M., Ren, Z., Wolock, C. J., Shneider, N. A., Phatnani, H., Cirulli, E. T., Lasseigne, B. N., Harris, T., et al. (2019), "A new approach for rare variation collapsing on functional protein domains implicates specific genic regions in ALS," *Genome research*, 29, 809–818.

Goeman, J. J. and Finos, L. (2012), "The inheritance procedure: multiple testing of tree-structured hypotheses," *Stat Appl Genet Mol Biol*, 11, Article 11.

Guo, W., Lynch, G., and Romano, J. P. (2018), "A new approach for large scale multiple testing with application to FDR control for graphically structured hypotheses," *arXiv preprint arXiv:1812.00258*.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The elements of statistical learning : data mining, inference, and prediction*, Springer.

Jenq, R. R., Ubeda, C., Taur, Y., Menezes, C. C., Khanin, R., Dudakov, J. A., Liu, C., West, M. L., Singer, N. V., Equinda, M. J., et al. (2012), "Regulation of intestinal inflammation by microbiota following allogeneic bone marrow transplantation," *Journal of Experimental Medicine*, 209, 903–911.

Jukes, T. H., Cantor, C. R., et al. (1969), "Evolution of protein molecules," *Mammalian protein metabolism*, 3, 21–132.

Kanoungi, G., Nothnagel, M., Becker, T., and Drichel, D. (2020), "The exhaustive genomic scan approach, with an application to rare-variant association analysis," *Eur J Hum Genet*, 28, 1283–1291.

Katsumata, Y. and Fardo, D. W. (2020), "Quantitative phenotype scan statistic (QPSS) reveals rare variant associations with Alzheimer's disease endophenotypes," *BMC Med Genet*, 21, 106.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015), "Sparse and compositionally robust inference of microbial ecological networks," *PLoS computational biology*, 11.

Lancaster, H. O. (1961), "Significance Tests in Discrete Distributions," *Journal of the American Statistical Association*, 56, 223–234.

Lee, D. and Lee, Y. (2016), "Extended likelihood approach to multiple testing with directional error control under a hidden Markov random field model," *Journal of Multivariate Analysis*, 151, 1 – 13.

Lee, S., Wu, M. C., and Lin, X. (2012), "Optimal tests for rare variant effects in sequencing association studies," *Biostatistics*, 13, 762–75.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014), "Rare-variant association analysis: study designs and statistical tests," *The American Journal of Human Genetics*, 95, 5–23.

Lesigne, E. (2005), *Heads or tails: An introduction to limit theorems in probability*, vol. 28, American Mathematical Soc.

Li, B. and Leal, S. M. (2008), "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," *Am J Hum Genet*, 83, 311–21.

Li, X., Sung, A., and Xie, J. (2021), "Distance Assisted Recursive Testing," *arXiv preprint arXiv:2103.11085*.

Li, Y., Hu, Y.-J., and Satten, G. A. (2020), "A Bottom-Up Approach to Testing Hypotheses That Have a Branching Tree Dependence Structure, With Error Rate Control," *Journal of the American Statistical Association*, pp. 1–18.

Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Zhou, H., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019), "Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies," *The American Journal of Human Genetics*, 104, 802–814.

Liu, J., Peissig, P., Zhang, C., Burnside, E., McCarty, C., and Page, D. (2012), "Graphical-model Based Multiple Testing under Dependence, with Applications to Genome-wide Association Studies," *Uncertain Artif Intell*, 2012, 511–522.

Liu, W. et al. (2013), "Gaussian graphical model estimation with false discovery rate control," *The Annals of Statistics*, 41, 2948–2978.

Liu, W., Shao, Q.-M., et al. (2014), "Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control," *The Annals of Statistics*, 42, 2003–2025.

Meijer, R. J. and Goeman, J. J. (2015), "A multiple testing method for hypotheses structured in a directed acyclic graph," *Biom J*, 57, 123–43.

Morgenthaler, S. and Thilly, W. G. (2007), "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST)," *Mutat Res*, 615, 28–56.

Morris, A. P. and Zeggini, E. (2010), "An evaluation of statistical approaches to rare variant analysis in genetic association studies," *Genet Epidemiol*, 34, 188–93.

Nam, S.-E., Cheung, Y. W. S., Nguyen, T. N., Gong, M., Chan, S., Lazarou, M., and Yip, C. K. (2021), "Insights on autophagosome–lysosome tethering from structural and biochemical characterization of human autophagy factor EPG5," *Communications biology*, 4, 1–14.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011), "Testing for an unusual distribution of rare variants," *PLoS Genet*, 7, e1001322.

Ng, P. C. and Henikoff, S. (2003), "SIFT: Predicting amino acid changes that affect protein function," *Nucleic acids research*, 31, 3812–3814.

Pan, W. (2009), "Asymptotic tests of association with multiple SNPs in linkage disequilibrium," *Genet Epidemiol*, 33, 497–507.

Pesiridis, G. S., Lee, V. M.-Y., and Trojanowski, J. Q. (2009), "Mutations in TDP-43 link glycine-rich domain functions to amyotrophic lateral sclerosis," *Human molecular genetics*, 18, R156–R162.

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. (2013), "Genic intolerance to functional variation and the interpretation of personal genomes," *PLoS genetics*, 9, e1003709.

Petrovski, S., Todd, J. L., Durheim, M. T., Wang, Q., Chien, J. W., Kelly, F. L., Frankel, C., Mebane, C. M., Ren, Z., Bridgers, J., et al. (2017), "An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis," *American journal of respiratory and critical care medicine*, 196, 82–93.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010), "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome research*, 20, 110–121.

Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A. S., and Goldstein, D. B. (2019), "Rare-variant collapsing analyses for complex traits: guidelines and applications," *Nature Reviews Genetics*, 20, 747–759.

Pura, J., Li, X., Chan, C., and Xie, J. (2019), "TEAM: A Multiple Testing Algorithm on the Aggregation Tree for Flow Cytometry Analysis," *arXiv preprint arXiv:1906.07757*.

Pura, J. A. (2020), "Multiple Testing Embedded in an Aggregation Tree With Applications to Omics Data," Ph.D. thesis, Duke University.

Roederer, M. and Hardy, R. R. (2001), "Frequency difference gating: a multivariate method for identifying subsets that differ between samples," *Cytometry Part A*, 45, 56–64.

Roederer, M., Moore, W., Treister, A., Hardy, R. R., and Herzenberg, L. A. (2001a), "Probability binning comparison: a metric for quantitating multivariate distribution differences," *Cytometry: The Journal of the International Society for Analytical Cytology*, 45, 47–55.

Roederer, M., Treister, A., Moore, W., and Herzenberg, L. A. (2001b), "Probability binning comparison: a metric for quantitating univariate distribution differences," *Cytometry: The Journal of the International Society for Analytical Cytology*, 45, 37–46.

Schliep, K. (2011), "phangorn: phylogenetic analysis in R," *Bioinformatics*, 27, 592–593.

Seder, R. A., Darrah, P. A., and Roederer, M. (2008), "T-cell quality in memory and protection: implications for vaccine design," *Nat Rev Immunol*, 8, 247–58.

Shu, H., Nan, B., and Koeppe, R. (2015), "Multiple testing for neuroimaging via hidden Markov random field," *Biometrics*, 71, 741–750.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005), "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome research*, 15, 1034–1050.

Soriano, J. and Ma, L. (2017), "Probabilistic multi-resolution scanning for two-sample differences," *Journal of The Royal Statistical Society Series B-statistical Methodology*, 79, 547–572.

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

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011), "Rare-variant association testing for sequencing data with the sequence kernel association test," *The American Journal of Human Genetics*, 89, 82–93.

Xia, Y., Cai, T. T., and Sun, W. (2019), "Gap: A general framework for information pooling in two-sample sparse inference," *Journal of the American Statistical Association*, pp. 1–15.

Xie, J. and Li, R. (2018), "False discovery rate control for high dimensional networks of quantile associations conditioning on covariates," *J R Stat Soc Series B Stat Methodol*, 80, 1015–1034.

Yekutieli, D. (2008), "Hierarchical False Discovery Rate-Controlling Methodology," *Journal of the American Statistical Association*, 103, 309–316.

Zhang, C., Fan, J., and Yu, T. (2011), "Multiple testing via FDR_L for large scale imaging data," *Annals of statistics*, 39, 613.

Zhao, Y. G., Zhao, H., Sun, H., and Zhang, H. (2013), "Role of Epg5 in selective neurodegeneration and Vici syndrome," *Autophagy*, 9, 1258–1262.