UMBC

DISSERTATION APPROVAL SHEET

Title of Dissertation: A Group Sequential Multiple Testing Method and Its Application to Genomic Data

Name of Candidate: Yewon Kim

Doctor of Philosophy, 2022

Graduate Program: Statistics

Dissertation and Abstract Approved:

Seungchul Baek

Seungchul Baek

Assistant Professor-Committee Chairperson

Department of Mathematics and Statistics

4/25/2022 | 1:42:08 PM EDT

Junyong Park

Junyong Park

Professor

Department of Statistics, Seoul National Univ

4/25/2022 | 5:05:38 오후 EDT

NOTE: *The Approval Sheet with the original signature must accompany the thesis or dissertation. No terminal punctuation is to be used.

ABSTRACT

Title of dissertation: A GROUP SEQUENTIAL MULTIPLE TESTING METHOD AND ITS APPLICATION TO GENOMIC DATA

Yewon Kim, Doctor of Philosophy, 2022

Dissertation directed by: Dr. Seungchul Baek
Department of Mathematics and Statistics
University of Maryland, Baltimore County

Dr. Junyong Park
Department of Statistics
Seoul National University

In this dissertation, we consider the simultaneous testing of groups and hypotheses within the groups which occurs in many scientific problems. A group is commonly judged to be significant if at least one hypothesis within the group is significant which is implemented via a global test for complete null hypothesis. However, this null hypothesis for group significance is strict, so all groups tend to be rejected especially when the number of hypotheses within a group is large. To avoid such trivial hypothesis testing results, we introduce the concept of margin to multiple testing problems so that we can adjust different levels of significance of the group. Based on this idea, we propose a group sequential multiple testing method with controlling false discovery rate (FDR) which incorporates the margin for group significance.

As real data applications, we apply the proposed method to functional groups

of single nucleotide polymorphisms (SNPs). We select significantly associated pairs of the summary statistics from genome-wide association study (GWAS) and linkage disequilibrium (LD) score. We further investigate additional local associations within haplotype blocks while existing methods such as LD score regression (LDSC) uses the whole SNPs. Our findings provide different aspects of explanation on the associations between the summary statistics and LD score such as Simpson's paradox.

In the second real data applications, we consider non-coding GWAS SNPs of regulatory DNA marked by deoxyribonuclease I (DNase I) hypersensitive sites (DHSs). By partitioning the GWAS SNPs for type 2 diabetes into DHSs groups, we apply the proposed method to detect statistically associated DHSs groups with type 2 diabetes. Each of the 32 DHSs groups represents a unique organ, the group related to the pancreas is detected as a significant group even with a large margin, and the findings are consistent with the intuition and published articles.

Some possible extensions of the proposed method and a summarization are discussed at the end of this dissertation.

A GROUP SEQUENTIAL MULTIPLE TESTING METHOD AND ITS APPLICATION TO GENETIC DATA

by

Yewon Kim

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Dr. Seungchul Baek, Chair/Advisor
Dr. Junyong Park, Co-Advisor
Dr. Anindya Roy
Dr. Dongwon Lee
Dr. Yaakov Malinovsky
Dr. Yi Huang

Acknowledgments

I sincerely appreciate my committee members, especially my advisors Dr. Seungchul Baek and Dr. Junyong Park for their endless support, kind and understanding during my graduate experience. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life.

I would also like to acknowledge with much appreciation the staff and faculty members at the UMBC Department of Mathematics and Statistics, who rendered their help during the period of my project work. It is their kind help and support that have made my study and life a wonderful time.

To friends and colleagues at UMBC, and my family, I express my sincere thanks. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

Table of Contents

List of Tables

List of Figures

Chapter 1: INTRODUCTION

1.1 Motivation

In fields such as biomedical science and genomics, modern statistical inference processes commonly involve testing a large number of hypotheses at the same time. For example, Laird and Lange (2010) mentioned that statistical genetics has played an important role in testing a relationship between genetic locations or genes and human diseases simultaneously. These genetic locations and genes often have a grouped structure for spatial, functional, biological or experimental process reasons. For these group structures, it is necessary to select important groups and genetic locations or genes within the groups. In the selection of significant groups, a global test for testing the complete null hypothesis is commonly used where the complete null hypothesis means all hypotheses within the group are null, so one should select a group as an important one as long as at least one null hypothesis within the group is rejected. This procedure may lead to some trivial result that all groups are declared to be significant in practice. One main reason for this result is that the null hypothesis for group importance is stringent, so the null hypothesis is too easily rejected only with a few significant hypotheses within the group. This phenomenon may occur in genetic and genomic studies since the number of groups and hypotheses

within the group are in general large in modern biological studies, for example, see Storey et al. (2007) and Laird and Lange (2010). We adjust the null hypothesis for group importance by requesting more significant hypotheses within the group by introducing the concept of the margin in the null hypothesis for group importance. The concept of margin is commonly used in clinical trials. For example, Althunian et al. (2017) mentioned that non-inferiority tests are used to check whether a new treatment is not worse than active treatment by more than a non-inferiority margin. See other related literature on the concept of margin such as Wiens (2002), Berger and Delampady (1987), Choi and Park (2014) and Chen and Chen (1999). Such a predetermined number of hypotheses within the group is the margin in this case which can be decided based on scientific knowledge or researcher's interest. We also expect sparsity between groups for group selection by controlling the margin. The results of changing the group significance criteria by the margin in real data applications will be discussed in Chapters 4 and 5.

1.2 Overview of Multiple Testing Procedures

Suppose that we have n null hypotheses $H_{0,1}, \cdots, H_{0,n}$ and n alternative hypotheses $H_{1,1}, \cdots, H_{1,n}$ to be tested simultaneously. The goal of a multiple testing procedure is developing a random or data driven rule for deciding which null hypotheses should be rejected or declared false. The decisions to reject or not the null hypotheses are based on the joint distribution of test statistics or p-values associated with the null hypotheses. The results of multiple hypothesis testings are

summarized in Table 1.1. Note that the total number of hypotheses n is fixed and known, and the number of true and false null hypotheses h_0 and h_1, respectively, are fixed and unknown. R_n is the number of rejected null hypotheses so it is observable, while W_n, U_n, V_n and S_n are not observable.

	true H_0	true H_1	
not reject H_0	W_n	U_n	$n - R_n$
reject H_0	V_n	S_n	R_n
	h_0	h_1	n

Table 1.1: Results of multiple hypothesis tests

Errors in hypothesis testing can be classified into type-I errors and type-II errors. Type-I error is the wrong rejection of a true null hypothesis, known as a false positive, while a type-II error is the incorrect acceptance of a false null hypothesis, known as a false negative. In Table 1.1, the chances of committing these two types of errors are inversely proportional: that is, decreasing type-I error increases type-II error and vice versa. In general, type-I errors are considered more harmful than type-II errors, and thus researchers bound the probability of making a type-I error by α, which represents an acceptable level of risk. The probability of making a type-I error is represented by α level. Using a lower value for α indicates that researchers will be less likely to detect a true alternative hypothesis, so risk of a type-II error will increase. The probability of having a type-II error is represented by β, and this is related to the power of the statistical test expressed as $1 - \beta$.

Sources of multiplicity arise in cases where researchers consider a set of statistical inferences simultaneously. In other words, the more statistical inferences are made, the probability of getting significant results simply by chance increases. Extensive research has been conducted on multiple testing procedures to address those problems, and we will review these errors and multiplicity correction in multiple hypotheses testing in the following sections.

1.2.1 Type-I Error Rates Based on the Distribution of the Number of Type-I Errors

We will review frequently used measurements of type-I error which are a function of only V_n.

- Family-wise error rate (FWER) is the probability of making at least one false rejection in a n series of hypothesis tests introduced by Tukey (1953), i.e.,

$$P(V_n > 0).$$

  For many scientific applications, n is particularly large, so the researcher can improve the ability of the procedure to detect false null hypotheses if one or more false rejections are tolerated, and the total number of false positive cases is controlled for k. Utilizing this idea, Lehmann and Romano (2005) proposed to replace the control of the FWER by controlling the false rejection probability over k, which is called generalized FWER.

- Generalized FWER (gFWER) is

$$P(V_n > k),$$

where k is predetermined.

- Per-family error rate (PFER) is the expected number of false rejections, i.e.,

$$E(V_n).$$

1.2.2 Type-I Error Rates Based on the Distribution of the Proportion of Type-I Errors Among the Rejected Hypotheses

We will review frequently used measurements of type-I error which are a function of V_n and R_n.

- False discover rate (FDR) is the expected proportion of type-I errors among the rejected hypotheses introduced by Benjamini and Hochberg (1995), i.e.,

$$E\Big(\frac{V_n}{R_n}\Big) = E\Big(\frac{V_n}{R_n}\Big|R_n > 0\Big)P(R_n > 0).$$

Note that a complete null hypothesis is when all n individual null hypotheses are true. Under the complete null hypothesis, FDR is equal to FWER. In general, R_n is greater than V_n, so FDR is less than or equal to FWER.

- Marginal FDR (mFDR) is the ratio of expected number of false discovery to the expected total number of rejected hypotheses, i.e.,

$$\frac{E(V_n)}{E(R_n)}.$$

- Storey (2002) considered a different definition of FDR, called for positive FDR, by conditioning on the cases that $R_n > 0$, i.e.

$$E\Big(\frac{V_n}{R_n}\Big) = E\Big(\frac{V_n}{R_n}\Big| R_n > 0\Big).$$

Note that the condition of $P(R_n > 0) = 1$ is common in most genomics experiments, where the FDR and positive FDR are very similar. Storey (2002) also discussed the advantages and disadvantages of positive FDR.

Another approach to use the concept of FDR to measure the statistical significance in genomic studies is the q-value introduced by Storey and Tibshirani (2003). q-value is similar to the existing p-value. In particular, p-value measures significance using false positive rate, while q-value utilizes the FDR. The local q-value is defined as the posterior probability that the null hypothesis is true given the p-value of the test, which is the same as the definition of the local FDR. Efron (2012) showed that the q-value and Benjamini and Hochberg (1995) procedures are equivalent. A detailed description of these procedures will be provided in the following Chapter 2.

FWER is appropriate when the researchers want to protect the results against any type-I errors. However, given the large number of hypotheses which is common in modern science, the researchers are often more interested in an efficient measure such as FDR. In this dissertation, we will use FDR as a measurement of overall type-I error rates.

1.2.3 Type-II Error Rates and Power

Statistical power in a hypothesis test is the probability that the test will correctly reject the null hypothesis when the alternative hypothesis is true, i.e.,

$$\text{power} = P(\text{reject } H_0 \mid H_1 \text{ is true}),$$

which represents true positive. Statistical power ranges from 0 to 1, and various concepts of power with an overall measure of type-II errors have been used in the literature. For example, as an analogue of FDR in terms of type-II errors, Genovese and Wasserman (2002) introduced false non-discovery rate (FNR), which is expected proportion of false non-discovery among the non-rejected null hypotheses, i.e.,

$$E\Big(\frac{U_n}{n-R_n}\Big) = E\Big(\frac{U_n}{n-R_n}\Big|n-R_n>0\Big)P(n-R_n>0).$$

1.3 Single Step Multiple Testing Procedures

We usually consider two main classes of multiple testing procedures, single step and multiple step or sequential procedures, depending on whether the thresholds for each test statistic or p-value leading to rejection of null hypothesis are random or constant. Specifically, in single step multiple testing procedures, each null hypothesis is tested using a threshold that is independent of the results of the tests of other hypotheses, and the threshold is not a function of the data. For example, the single step multiple testing procedure uses equal adjustments to each p-value in simultaneous n series hypotheses testing. This procedure keeps the overall type-I errors at the desired α level, which is called the Bonferroni correction (Bonferroni,

1936). The Bonferroni procedure can be used to control the FWER. In order to utilize Bonferroni bound, we can divide the target α level by the total number of hypothesis tests and apply the updated α level to each individual hypothesis test for finding significant hypothesis. For any p-value which is less than the updated α value, the corresponding null hypothesis is rejected.

Although the Bonferroni bound can be utilized to control the FWER, this bound has been criticized for small power and a high probability of type-II errors. Improvement in power, while preserving type-I error control, may be achieved by multiple step multiple testing procedures, in which the decision to reject a particular null hypothesis depends on the outcome of the tests of other hypotheses. In other words, the multiple step multiple testing procedures use an adaptive adjustment for each p-value in the n series hypothesis testing.

1.4 Multiple Step Multiple Testing Procedure

Multiple step multiple testing procedures are of two main types, step down and step up procedures, depending on the order in which the null hypotheses are tested.

In step down multiple testing procedures, the null hypothesis with the smallest p-value is tested first. In particular, as long as one of null hypotheses fails to be rejected, all the hypotheses with larger p-values would fail to be rejected. For example, Holm (1979) controlled FWER for arbitrary dependent p-value structures and this method provided higher statistical power than the Bonferroni correction.

On the other hand, the step up multiple testing procedure starts from the largest p-value and rejects all smaller p-values after the first one is rejected. For example, Hochberg (1988) developed to control FWER.

Step down multiple testing procedures are known to be more conservative than step up multiple testing procedures. Therefore, in order to control for type-I error rates, the step up multiple testing procedures are subject to more conditions than the step down multiple testing procedures. FWER control employs a more stringent control over false discovery compared to FDR controlling procedures. FDR controlling procedures have higher statistical power at the cost of increased rates of type-I errors. Literature with multiple step simultaneous testing procedures and multiple testing methods for controlling FDR will be discussed in the following Chapter 2.

Chapter 2: LITERATURE REVIEW

In this chapter, we will review several FDR controlling methods and simultaneous grouped hypothesis testing with controlling FDR that are relevant to the methodology discussed in this dissertation.

2.1 FDR Controlling Approaches

2.1.1 Benjamini and Hochberg (1995) Procedure

Benjamini and Hochberg (1995) introduced a step up multiple testing procedure to control FDR, which is referred to as the BH procedure in the literature. Consider n null hypotheses $H_{0,1}, \cdots, H_{0,n}$ and the corresponding p-values $p_1, \cdots, p_n$. In particular, $p_{(1)} \leq \cdots \leq p_{(n)}$ denote the ordered p-values and $H_{0,(i)}$ being the null hypothesis corresponding to $p_{(i)}$. The BH procedure reject k null hypotheses such that

$$k = max\left\{ m : p_{(m)} \leq \frac{m}{n}\alpha \right\}.$$

Benjamini and Hochberg (1995) proved that if all p-values are independent, rejecting $H_{0,(1)}, \cdots, H_{0,(k)}$ provides the FDR at the target level α. Furthermore, Benjamini and Yekutieli (2001) proved that the BH procedure also controls the FDR when

the test statistics have positive regression dependence on subset (PRDS) on each of the test statistics corresponding to the true null hypotheses. For the arbitrary dependency cases, Benjamini and Yekutieli (2001) showed that the FDR can be controlled at the target level α with a conservative and simple modification of the BH procedure.

The FDR depends on the overall proportion of null hypothesis denoted by π_0, and Storey (2002) mentioned a more powerful multiple testing procedure compared to the BH method, which used information of π_0 when estimating the FDR. The method of Storey (2002) reject k null hypotheses such that

$$k = max\left\{ m : p_{(m)} \leq \frac{1}{\widehat{\pi_0}} \frac{m}{n} \alpha \right\},$$

where $\hat{\pi_0}$ is an estimator of π_0. Storey (2002) also noted that the power of the multiple testing approach does not necessarily decrease when more hypotheses are considered, as the estimator of π_0 improves when the large number of hypotheses are simultaneously tested. One of the important ideas of the methodology of Storey (2002) is to use data information to get better estimates of π_0 to improve the performance of multiple testing procedures. In this dissertation, we will utilize group information to estimate π_0 for each group in the multiple testing methodology which we will discuss in the Chapter 3.

2.1.2 Empirical Bayes Procedure

In order to detect genes affected by radiation treatment among 700 human genes, Efron et al. (2001) introduced a non-parametric empirical Bayes model with

local FDR which is closely related FDR. Consider the random mixture density for activity of gene $z \sim f(z) = (1 - \pi_0)f_1(z) + \pi_0 f_0(z)$ where $f_0(z)$ is a null density and $f_1(z)$ is a non null density. Here, π_0 is an unknown mixing probability. With an application of Bayes theorem to the mixture model, Efron et al. (2001) defined local FDR as

$$\text{local FDR} = \frac{\pi_0 f_0(z)}{f(z)},$$

which is the a posterior probability of the unaffected by radiation treatment class given z. Furthermore, Efron et al. (2001) estimated the mixture density using Poisson regression using data. For the estimation of π_0 and f_0, Efron et al. (2001) used zero assumption, as the condition that most of the probability mass near the mode of f is from the null density. Standard normal distribution is another candidate for the null distribution, and the further discussion of choosing a null distribution is covered in Efron (2004).

2.2 Grouped Multiple Testing Methodology

Large-scale genomic data approaches have enabled us to analyze genome-wide methylation patterns and to enumerate DNA sequence alterations across the genome. In some genomic data, the data can be grouped by functionally, structurally and spatially, in which case ignoring group structure in data analysis can be misleading the results of research. When the researchers can define such groups, it may be desirable to control the group-wise FDR or local FDR simultaneously for all groups. In this section, we will review some existing group structured multiple

testing procedures with controlling FDR.

2.2.1 Heller et al. (2018) Method

Heller et al. (2018) considered multiple testing problems of grouped hypotheses using gene expression data, namely post-selection inference. They performed FWER or FDR controlling procedures to find significant gene sets or groups. In particular, for selecting significant groups, Heller et al. (2018) used the global null hypothesis. With the selected gene sets, Heller et al. (2018) performed the conditional FWER or FDR controlling procedures in order to identify significantly differentially expressed genes within the selected gene sets. Under specific model assumptions, Heller et al. (2018) proved that the BH method at level α controls the conditional FDR at level $\frac{n_0}{n}\alpha$ where n_0 is the number of true null hypotheses and n is the number of hypotheses within the gene groups. In this case, n_0 is an unknown quantity, however $\frac{n_0}{n}$ is less than or equal to 1. The condition of $\frac{n_0}{n}\alpha \leq \alpha$ means this methodology is conservative.

2.2.2 Sarkar et al. (2019) Method

Sarkar et al. (2019) considered n hypotheses which were simultaneously tested in a k-group sequential method. Under positively dependent through stochastic ordering (PDS) assumption of p-values, Sarkar et al. (2019) proposed the k-stage group sequential BH method, and proved the FDR of the process is controlled by $\pi_0\alpha$ where π_0 is the null proportion. When the p-values are independent across

hypotheses and PDS assumption holds, Sarkar et al. (2019) extended the method to adaptive group sequential BH method. The extended model updates the π_0 estimation at each step of the algorithm. Sarkar et al. (2019) showed numerical validation of the extended method's FDR controlling while examining the extended method's performance relative to other competitors.

2.2.3 Barber and Ramdas (2016) Method

Barber and Ramdas (2016) suggested multiple testing methods for group-FDR based on arbitrary partitioned p-values. The method of Barber and Ramdas (2016) considered a list of n p-values and handled all non-hierarchical partitions. For example, with the finest partition into n singletons, the result of this method is consistent with the result of the BH method. For a single group of size n, the result of this method is the same as Simes test for the global null hypothesis. Barber and Ramdas (2016) showed the results of methodology with grouped hypotheses according to brain regions using fMRI data.

2.2.4 Liu et al. (2016) Method

Liu et al. (2016) provided a group sequential multiple testing process to control FDR. In particular, it considered group level FDR based on several groups by expanding multiple testing procedures for a single group. By considering the within-group local FDR for each group and group-wise posterior probability for group significance, this method integrated the process of finding significant hypotheses within

each group and the importance of that group to find hypotheses that can finally be discovered. For a group to be significant, a condition is required that at least one hypothesis is significant within the group. Liu et al. (2016) showed that if only a few groups are actually significant in small-scale to medium-scale data, this methodology is more effective than other methodologies that ignore the group structure under certain model conditions.

2.2.5 Sun et al. (2015) Method

Sun et al. (2015) considered cluster-wise multiple testing inference of spatial signals using marginal FDR. Using the monotone ratio condition (MRC), Sun et al. (2015) defined false cluster-wise discovery rate (FCR) and marginal FCR. For the measurement of power, Sun et al. (2015) introduced missed discovery rate (MDR). Based on real data application and simulation study, Sun et al. (2015) showed that the proposed method can control marginal FCR as well as asymptotically control FCR under some condition. However, in some cases the proposed method failed to control FCR as proved by Heller et al. (2018). This is because marginal FDR is more conservative than FDR.

Chapter 3: THE PROPOSED METHOD AND SIMULATION STUDIES

In this chapter, we present a proposed multiple testing method for grouped hypotheses controlling FDR. The proposed methodology is more efficient when the number of hypotheses within a group is large, and it includes the case of global testing. Between and within groups hypotheses are evaluated for significance using the posterior probability, and only significant hypotheses within a significant group can be discovered. We examine the effectiveness of the proposed methodology through simulations when the number of hypotheses in the group is all the same as well as when the number of hypotheses is unbalanced.

3.1 The Proposed Methodology

There are two hypotheses that we are interested in here, one for the significance of individual hypotheses and the other for group significance. As a criterion for judging the significance of a group, a popular criterion is that the group is considered significant if at least one hypothesis in the group is significant. If the number of hypotheses in a group is large, the complete null hypothesis for group significance is too strict, so the finding of a meaningful group is too trivial in the sense that all

groups are easily considered as significant. In this case, the selection of a significance group under the complete null hypothesis does not provide any information on group selection when all groups are declared to be significant. Instead, we give a more flexible null hypothesis for group significance which depends on the number of non-null hypotheses within the group. In other words, if the number of significant hypotheses within the group is not enough, the group itself becomes a non-significant group, and if the number of non-null hypotheses within a group is enough, the group itself becomes a significant group. We call a thresholding value of the number of null hypotheses within the group called the margin of the null group. The hypothesis reflecting this idea and the corresponding FDR controlling procedure are presented in the following sections.

3.1.1 Hypotheses with Margin

Consider the group structured data x_{gj} corresponding to the jth hypothesis in the gth group for $g = 1, 2, \cdots, G$, $j = 1, 2, \cdots, m_g$, and $\mathbf{x}_g = (x_{g1}, x_{g2}, \cdots, x_{gm_g})^T$. Let binary random variable θ_g be the indicator of the significance of the gth group where $\theta_g = 1$ if gth group is significant and $\theta_g = 0$ otherwise, i.e., the hypotheses on the significance of groups are

$$H_{0,g} : \theta_g = 0 \quad vs. \quad H_{1,g} : \theta_g = 1,$$

for $g = 1, \ldots, G$. We also define $\boldsymbol{\theta}_g = (\theta_1, \theta_2, \cdots, \theta_G)^T$ which is used as an overall group-wise significance test.

Given group membership such as $\theta_g = 0$ or 1, the state for the jth hypothesis

in the gth group is denoted by $\theta_{j|g}$ where $\theta_{j|g}$ has a value of 0 if the corresponding null hypothesis is true and 1 otherwise. Testing of hypotheses within the group using the given group condition are

$$H_{0,j|g} : \theta_{j|g} = 0 \;\; vs. \;\; H_{1,j|g} : \theta_{j|g} = 1,$$

for $g = 1, 2, \cdots, G$, $j = 1, 2, \cdots, m_g$.

Let $\boldsymbol{\theta}_{j|g} = (\theta_{1|g}, \theta_{2|g}, \cdots, \theta_{m_g|g})^T$ be the vector indicating either the true null or alternative within gth group. On the other hand, the decision rules are denoted by $\delta_g(\mathbf{x}_g) \in \{0, 1\}$ and $\delta_{j|g}(\mathbf{x}_g) \in \{0, 1\}$ corresponding to θ_g and $\theta_{j|g}$ respectively, where 1 means to reject the corresponding null hypothesis and 0 otherwise. Consider the following hierarchical models for θ_g, $\theta_{j|g}$ and x_{gj}:

$$\theta_1, \theta_2, \cdots, \theta_G \overset{\text{i.i.d}}{\sim} Bernoulli(\pi_1), \tag{3.1}$$

$$\theta_{1|g}, \theta_{2|g}, \cdots, \theta_{m_g|g} | \theta_g = 0 \sim Bernoulli(\epsilon_g) I\bigg(\sum_{j=1}^{m_g} \theta_{j|g} \leq M_g \bigg), \tag{3.2}$$

$$\theta_{1|g}, \theta_{2|g}, \cdots, \theta_{m_g|g} | \theta_g = 1 \sim Bernoulli(\epsilon_g) I\bigg(\sum_{j=1}^{m_g} \theta_{j|g} > M_g \bigg), \tag{3.3}$$

$$x_{gj} | \theta_{j|g} \sim f_g(x_{gj}) = (1 - \theta_{j|g}) f_{0,g}(x_{gj}) + \theta_{j|g} f_{1,g}(x_{gj}), \tag{3.4}$$

where M_g is the threshold that determines the membership of the gth group. $f_{0,g}$ and $f_{1,g}$ are the densities of the null and non-null distributions for the gth group. If M_g is 0, testing the significance of a group becomes a testing problem for the complete null hypothesis that $\bigcap_{j=1}^{m_g} H_{0,j|g}$ are true. In the case of $M_g > 0$, the importance of the group is concluded that the group is meaningful only when the number of non-null hypotheses in the group is more than the appropriate given M_g. In our context, the significance of the gth group is based on giving the margin or tolerance to the null

hypothesis through $\sum_{j=1}^{m_g} \theta_{j|g} \leq M_g$ rather than setting $M_g = 0$. As M_g increases, the decision of the significance of the gth group is getting more conservative, since the condition requests stronger signals in the gth group. From this formulation, we test

$$H_{0,gj} : \theta_{gj} = 0 \;\; vs. \;\; H_{1,gj} : \theta_{gj} = 1,$$

where $\theta_{gj} = \theta_{j|g}\theta_g$. We see that the jth null hypothesis in the gth group is rejected when both $\theta_{j|g} = 1$ and $\theta_g = 1$. In other words, in order for a particular hypothesis to be significant, the group containing the hypothesis must also be significant, so testing the significance of groups is followed by testing the jth hypothesis in the gth group. In the following section, we present the FDR controlling procedure incorporating (3.1)-(3.4).

3.1.2 Estimation of Null Distributions and Parameters

For the case of $M_g = 0$, Liu et al. (2016) provided a two-fold loop testing algorithm (TLTA) which is a multiple testing procedure of group structured data to control FDR at a given level of α. To illustrate the TLTA, Liu et al. (2016) used $PFDR_T(\mathbf{x})$, the expected false discovery proportion conditional on $\mathbf{x} = \{x_{gj}\}$ for all (g, j), and shows that the multiple testing procedure based on $PFDR_T(\mathbf{x})$ controls a given level of FDR where

$$PFDR_T(\mathbf{x}) = E\left(\frac{\sum_{g=1}^{G} \sum_{j=1}^{m_g} (1 - \theta_{gj}) \delta_{gj}(\mathbf{x}_g)}{\{\sum_{g=1}^{G} \sum_{j=1}^{m_g} \delta_{gj}(\mathbf{x}_g)\} \vee 1} \middle| \mathbf{x} \right) \leq \alpha, \tag{3.5}$$

for $\delta_{gj}(\mathbf{x}_g) = \delta_g(\mathbf{x}_g)\delta_{j|g}(\mathbf{x}_g)$ and $a \vee b = \max(a, b)$. We extend this procedure to the case that the hypotheses of groups have a margin as described in (3.2) and

(3.3). The likelihood function of $\mathbf{\Phi} = (\pi_1, \epsilon_g)$ given data $\mathbf{x}$ and latent variables $\mathbf{\Theta} = (\boldsymbol{\theta}_g, \boldsymbol{\theta}_{j|g})$ for all (g, j) is

$$L(\mathbf{\Phi}|\mathbf{x}, \mathbf{\Theta}) = \prod_{g=1}^{G} \bigg(\pi_1 P(\mathbf{x}_g|\theta_g = 1) + (1 - \pi_1) P(\mathbf{x}_g|\theta_g = 0) \bigg), \tag{3.6}$$

where

$$P(\mathbf{x}_g|\, \theta_g = 1) = \sum_{\boldsymbol{\theta}_{j|g} \in \Omega_{g1}} \left(\prod_{j=1}^{m_g} \frac{\epsilon_g^{\theta_{j|g}} (1 - \epsilon_g)^{1-\theta_{j|g}}}{P(\sum_{k=1}^{m_g} \theta_{k|g} > M_g)} f_{\theta_{j|g}}(x_{gj}) \right),$$
$$P(\mathbf{x}_g|\, \theta_g = 0) = \sum_{\boldsymbol{\theta}_{j|g} \in \Omega_{g0}} \left(\prod_{j=1}^{m_g} \frac{\epsilon_g^{\theta_{j|g}} (1 - \epsilon_g)^{1-\theta_{j|g}}}{P(\sum_{k=1}^{m_g} \theta_{k|g} \leq M_g)} f_{\theta_{j|g}}(x_{gj}) \right),$$

for $\Omega_{g0} = \{\boldsymbol{\theta}_{j|g} : \sum_{j=1}^{m_g} \theta_{j|g} \leq M_g\}$ and $\Omega_{g1} = \{\boldsymbol{\theta}_{j|g} : \sum_{j=1}^{m_g} \theta_{j|g} > M_g\}$. More detailed derivation is given in Appendix A.1.

Since $f_{\theta_{j|g}}$ for $\theta_{j|g} = 0$ or 1 is the mixture of $f_{0,g}$ and $f_{1,g}$, we need to identify them. It is typical that the distribution $f_{0,g}$ is assumed to be known, such as a standard normal distribution, for example, Liu et al. (2016) used a standard normal distribution as a null distribution. However, in many practical problems, such a standard normal distribution may not reflect real situations such as correlated data and existence of covariates. To overcome these difficulties, Efron (2004) introduced the estimation of the null distribution $f_{0,g}$ under an empirical Bayes setting based on the center of the data. In this dissertation, the null distributions $f_{0,g}$ for $1 \leq g \leq G$ for different groups are estimated to avoid uncertainty from a standard normal distribution as a null distribution. See Efron (2004) for more detail. For gth group, we estimate ϵ_g, $f_{0,g}$ and f_g which can be obtained in `locfdr` in R-package,

$$\hat{f}_{0,g} \sim N(\hat{\delta}_g, \hat{\sigma}_g^2),$$

$\hat{\epsilon}_g$ and $\hat{f}_g$ where $\hat{f}_g$ is estimated using Poisson regression. When we have estimators $(\hat{\epsilon}_g, \hat{f}_{0,g}, \hat{f}_g)$, $1 \le g \le G$, we estimate π_1 using the EM algorithm (Dempster et al., 1977) as follows:

Algorithm 1 EM algorithm of π_1

1: Set an initial $\hat{\pi}_1^{(0)} = 1/2$.

2: Repeat the following steps until $\hat{\pi}_1^{(l)}$ converges:

$$\begin{aligned}\hat{\pi}_1^{(l)} &= \frac{1}{G}\sum_{g=1}^{G} P(\theta_g = 1|\mathbf{x}_g, \hat{\pi}_1^{(l-1)}) = \frac{1}{G}\sum_{g=1}^{G} \frac{\hat{\pi}_1^{(l-1)} P(\mathbf{x}_g|\theta_g = 1)}{g^{(l-1)}(\mathbf{x}_g)} \\ &= \frac{1}{G}\sum_{g=1}^{G} \frac{\hat{\pi}_1^{(l-1)} \dfrac{P(\sum_{k=1}^{m_g}\theta_{k|g} > M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g} > M_g)}}{\hat{\pi}_1^{(l-1)} \dfrac{P(\sum_{k=1}^{m_g}\theta_{k|g} > M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g} > M_g)} + (1-\hat{\pi}_1^{(l-1)}) \dfrac{P(\sum_{k=1}^{m_g}\theta_{k|g} \le M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g} \le M_g)}},\end{aligned}$$

where $g^{(l-1)}(\mathbf{x}_g) = \hat{\pi}_1^{(l-1)} P(\mathbf{x}_g|\theta_g = 1) + (1 - \hat{\pi}_1^{(l-1)}) P(\mathbf{x}_g|\theta_g = 0)$.

Detailed derivation of Algorithm 1 is given in Appendix A.2.

We need $P(\sum_{k=1}^{m_g}\theta_{k|g} > M_g|\,\mathbf{x}_g)$ and $P(\sum_{k=1}^{m_g}\theta_{k|g} > M_g)$ for each group to Algorithm 1, so we define the following probabilities and approximations. First of all, we consider $\widetilde{fdr}_{gj} = P(\theta_{gj} = 0|\mathbf{x}_g)$ as the within-group local FDR,

$$\widetilde{fdr}_{gj} = \frac{(1-\epsilon_g) f_{0,g}(x_{gj})}{f_g(x_{gj})}, \quad 1 - \widetilde{fdr}_{gj} = \frac{\epsilon_g f_{1,g}(x_{gj})}{f_g(x_{gj})}. \tag{3.7}$$

There are m_g tests in each group, and (3.7) is an essential factor in the within-group discoveries. Secondly, by the law of total probability, the normalizing constant of truncated Bernoulli distribution of $\boldsymbol{\theta}_{j|g}$ in (3.2) and (3.3) are

$$\begin{aligned}P(\sum_{j=1}^{m_g}\theta_{j|g} \le M_g) &= \sum_{\boldsymbol{\theta}_{j|g}\in\Omega_{g0}} \left(\prod_{j=1}^{m_g}(1-\epsilon_g)^{1-\theta_{j|g}}(\epsilon_g)^{\theta_{j|g}}\right), \\ P(\sum_{j=1}^{m_g}\theta_{j|g} > M_g) &= \sum_{\boldsymbol{\theta}_{j|g}\in\Omega_{g1}} \left(\prod_{j=1}^{m_g}(1-\epsilon_g)^{1-\theta_{j|g}}(\epsilon_g)^{\theta_{j|g}}\right),\end{aligned}$$

where $\Omega_{g0} = \{\boldsymbol{\theta}_{j|g} : \sum_{j=1}^{m_g} \theta_{j|g} \leq M_g\}$ and $\Omega_{g1} = \{\boldsymbol{\theta}_{j|g} : \sum_{j=1}^{m_g} \theta_{j|g} > M_g\}$. However, the probabilities above are computationally intractable at large m_g, $1 \leq g \leq G$. We apply the binomial-normal approximation with m_g and ϵ_g for the gth group

$$P(\sum_{j=1}^{m_g} \theta_{j|g} \leq M_g) \approx \Phi\left(\frac{M_g - m_g\epsilon_g}{\sqrt{m_g\epsilon_g(1-\epsilon_g)}}\right),$$
$$P(\sum_{j=1}^{m_g} \theta_{j|g} > M_g) \approx 1 - \Phi\left(\frac{M_g - m_g\epsilon_g}{\sqrt{m_g\epsilon_g(1-\epsilon_g)}}\right).$$

The probabilities above provide the prior probability of a group membership. Similarly, the posterior probability of a group membership uses the binomial-normal approximation for computational efficiency,

$$\sum_{k=1}^{m_g} \theta_{k|g}|\, \mathbf{x}_g \approx N\left(\sum_{k=1}^{m_g}(1 - \widetilde{fdr}_{gk})\,,\, \sum_{k=1}^{m_g} \widetilde{fdr}_{gk}(1 - \widetilde{fdr}_{gk})\right). \tag{3.8}$$

In practice, $(\epsilon_g, f_{0,g}, f_g)$, $1 \leq g \leq G$ are replaced by their estimators obtained in `locfdr` R-package. $\sum_{k=1}^{m_g} \theta_{k|g} \leq M_g|\mathbf{x}_g$ and $\sum_{k=1}^{m_g} \theta_{k|g} > M_g|\mathbf{x}_g$ can be expressed as a binomial distribution, considering m_g independent hypotheses and $(1 - \widetilde{fdr}_{gj})$ as success probability of the jth hypothesis in the gth group.

3.1.3 Proposed Multiple Testing Procedure

The control of $PFDR_T(\mathbf{x})$ guarantees the control of the FDR at level α. According to (3.5), we have

$$E_{\mathbf{x}}(PFDR_T(\mathbf{x})) = E_{\mathbf{x}}\left\{E\left(\frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}(1-\theta_{gj})\delta_{gj}(\mathbf{x})}{\{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\mathbf{x})\} \vee 1}\middle|\mathbf{x}\right)\right\} \leq \alpha. \tag{3.9}$$

The following theorem shows that $PFDR_T(\mathbf{x})$ is expressed based on $\delta_g(\mathbf{x}_g)$ and $\delta_{j|g}(\mathbf{x}_g)$ under the situation (3.1)-(3.4) which provides the main idea of a group

sequential multiple testing procedure with controlling FDR. Based on all estimators in the previous section and Theorem 1, we propose a multiple testing procedure in Algorithm 2.

Theorem 1 *Under the settings* (3.1)-(3.4)*, we have the* $PFDR_T(\mathbf{x})$ *defined in* (3.5) *as follows :*

$$PFDR_T(\mathbf{x}) = \frac{\sum_{g=1}^{G}\delta_g(\mathbf{x}_g)[1-w_g\{I(\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g))-PFDR_{w|g}(\mathbf{x}_g)\}]\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g)}{\sum_{g=1}^{G}\delta_g(\mathbf{x}_g)\{\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g)\}\vee 1},$$

where

$$w_g = \frac{\frac{\pi_1 P(\sum_{j=1}^{m_g}\theta_{j|g} > M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g} > M_g)}}{\frac{(1-\pi_1)P(\sum_{j=1}^{m_g}\theta_{j|g} \le M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g} \le M_g)} + \frac{\pi_1 P(\sum_{j=1}^{m_g}\theta_{j|g} > M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g} > M_g)}}$$

and

$$PFDR_{w|g}(\mathbf{x}_g) = \frac{\sum_{j=1}^{m_g}\widetilde{fdr}_{gj}\delta_{j|g}(\mathbf{x}_g)}{\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g)\vee 1}.$$

Proof 1 *See A.3 in Appendix.*

Based on Theorem 1, we propose the following multiple testing procedure which selects significant groups and hypotheses within the corresponding groups.

Algorithm 2 Proposed multiple testing procedure

1: The group membership is determined based on π_1 obtained from Algorithm 1.

2: For each group, let $\widetilde{fdr}_{g(1)} \le \widetilde{fdr}_{g(2)} \le \cdots \le \widetilde{fdr}_{g(m_g)}$ be the ordered $\widetilde{fdr}_{gj}$ with $H_{g(1)}, \cdots, H_{g(m_g)}$ being the corresponding null hypotheses. For the gth group, we reject R_g hypotheses, $H_{g(1)}, \ldots, H_{g(R_g)}$ where

$$R_g = max\left(k_g : \frac{1}{k_g}\sum_{j=1}^{k_g}\widetilde{fdr}_{g(j)} \le \eta\right), \tag{3.10}$$

given $0 \le \eta \le \alpha$, where η is a threshold to control the within-group local FDR.

3: For each group, compute $\eta_g = \frac{1}{R_g}\sum_{j=1}^{R_g}\widetilde{fdr}_{g(j)}$ and $fdr_g^* = 1 - w_g(1-\eta_g)$ where w_g is defined in Theorem 1.

4: For ordered $fdr_g^* : fdr_{(1)}^* \leq fdr_{(2)}^* \cdots \leq fdr_{(G)}^*$, find out

$$l = max\left(k : \frac{\sum_{g=1}^{k} R_{(g)} fdr_{(g)}^*}{\sum_{g=1}^{k} R_{(g)}} \leq \alpha\right), \tag{3.11}$$

where $R_{(g)}$ is the number of possible discoveries for the group corresponding to $fdr_{(g)}^*$.

5: Reject the first l groups corresponding to $fdr_{(i)}^*$ for $1 \leq i \leq l$ and the total $\sum_{i=1}^{l} R_{(i)}$ hypotheses within the rejected groups.

Remark 1 *An additional threshold η in* (3.10) *is employed to control the within-group local FDR for each group, which is corresponding to control $PFDR_{w|g}(\mathbf{x}_g)$ in Theorem 1. For example, the number of selected hypotheses in the group in* (3.10) *will decrease for small values of η. Due to the constraints of α in* (3.11)*, we observe a trade-off relationship between the number of significant groups and hypotheses within the group. As η increases for some given margin, the number of significant hypotheses within each group increases, while the number of significant groups decreases due to the constraint of α. This phenomenon will be discussed further in Section 3.2, Chapter 4 and 5.*

Remark 2 *Since FDR is regulated for all $\eta \leq \alpha$, we can choose η according to our purpose. The margin is predetermined based on some prior knowledge, if we focus more on the number of selected groups, we can choose η that maximizes the number of groups at a given margin. On the other hand, if we want to maximize the number of selected hypotheses, we can choose such η for a given margin. In other*

words, researchers may examine the number of significant groups and hypotheses for the grid values of $\eta \in (0, \alpha]$ at a given margin. Based on the obtained results, the researchers are able to select an appropriate η according to the purpose of the study. Algorithm 2 can control FDR even including the process of finding η.

3.2 Numerical Studies

Simulations to verify the performance of our algorithm are conducted in two cases. The first case deals with equal-sized groups, and the second case handles unequal group sizes. For all simulations, we set $G = 50$, $\alpha = 0.05$, $f_{0,g}(x) = \phi(x)$ and $f_{1,g}(x) = \phi(x - \mu)$, $2.5 \leq \mu \leq 3.5$ and $\phi(x)$ is a standard normal distribution. In Algorithm 2, η in (3.10) is a given value, so we use grid points of η from 0.02 to 0.05 to see the results corresponding to those different values of η. We set the number of non-significant groups to 40 and that of significant groups to 10, leading to sparsity between groups. The margin M_g in (3.2) and (3.3) is

$$M_g = cm_g,$$

where c is a given constant. Here, c controls the margin, hence we reach a conservative decision on the number of significant groups for larger values of c. In our simulation studies, we report simulations on grid points of c from 0 to 0.06. Given the samples generated from $f_{0,g}$ and $f_{1,g}$, we employ `locfdr` R-package to estimate ϵ_g, $f_{0,g}$, f_g and the within-group local FDR for each group. All simulation results are the mean of 100 independent simulations.

3.2.1 Simulation 1 : Equal Group Size

First, we set up all groups to have the equal sample sizes, $m_g = 10,000, 1 \leq g \leq G$. We report the results of simulations implemented under the model assumption (3.1)-(3.4) to evaluate the performance of Algorithm 2. In particular, power is defined as the expected proportion of alternative hypotheses that are correctly discovered. For the predetermined c and η, the simulation results are shown in Figure 3.1 - 3.4. When c is 0, $M_g = cm_g$ becomes 0. This is the case considered in Liu et al. (2016) which means the groups are selected based on the complete null hypothesis $\prod_{j=1}^{m_g} I(\theta_{j|g} = 0) = I(\sum_{j=1}^{m_g} \theta_{j|g} = 0)$. Specifically, Figure 3.1 - 3.4 shows the number of selected groups, the total number of selected hypotheses, statistical power, and controlling the FDR, respectively. According to Figure 3.1 - 3.2, it is observed that for a given value of $c \in [0, 0.02)$, there is a trade-off relationship between the number of significant groups and hypotheses within the groups as η changes, however this phenomenon does not occur all the time. For all values of c and η, Figure 3.4 shows that FDR is controlled by α level which is from (3.9).

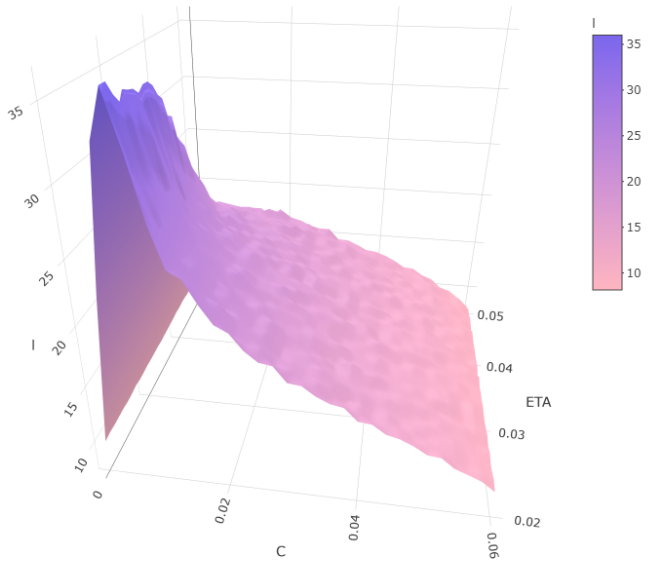

Figure 3.1: Simulation results of the number of significant groups when the group sizes are balanced

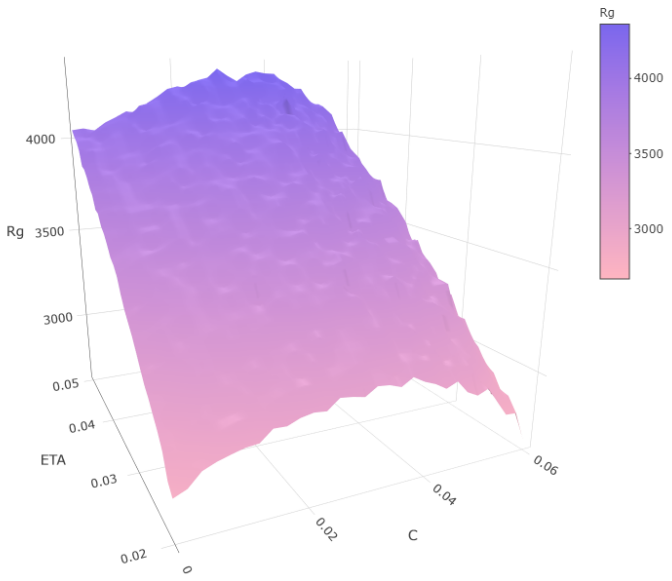

Figure 3.2: Simulation results of the total number of significant hypotheses when the group sizes are balanced

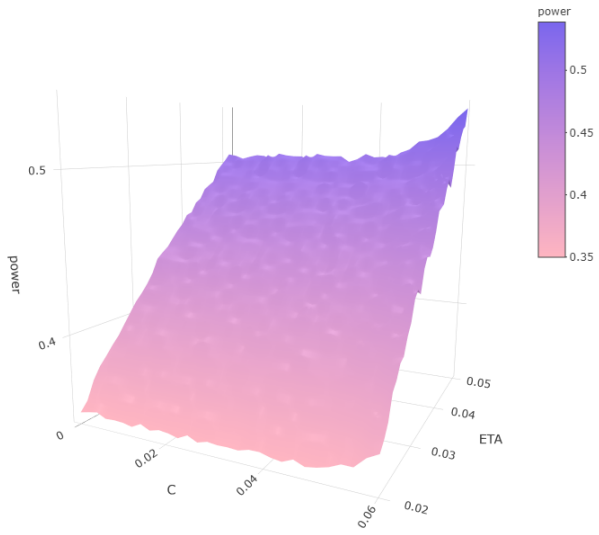

Figure 3.3: Simulation results of power when the group sizes are balanced

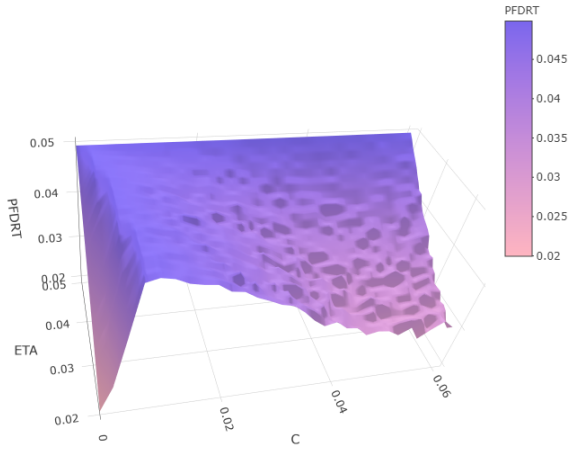

Figure 3.4: Simulation results of controlling the FDR when the group sizes are balanced

3.2.2 Simulation 2 : Unequal Group Size

In the second case of simulations, we consider that the group sizes are unbalanced. All other settings are the same as simulation 1. The summary table of group sizes is shown in Table 3.1 which mimics the real data example in Chapter 4 and 5. For the given values of c and η, the simulation results are represented in Figure 3.5 - 3.8 showing that the overall results are similar to those of simulation 1.

min	Q_1	median	mean	Q_3	max
8,046	11,138	28,546	59,486	64,049	420,432

Table 3.1: Summary table for the group sizes for 52 groups in unequal group size simulation study.

However, for all pairs of c and η, the number of rejected hypotheses in Figure 3.6 tends to be less smooth than simulation 1. For a given c, we observe that there is more dynamical change for each stimulation in the selected groups and hypotheses of the groups for different values of η. Accordingly, the sums of the selected hypotheses tend to have more variation than those from simulation 1. We also observe that the trade-off relationship exists between the number of significant groups and hypotheses within the groups for any given $c \in [0, 0.03)$ depending on η. Similar to simulation 1, Figure 3.8 shows that the FDR in all the scenarios are controlled at the level α by (3.9).

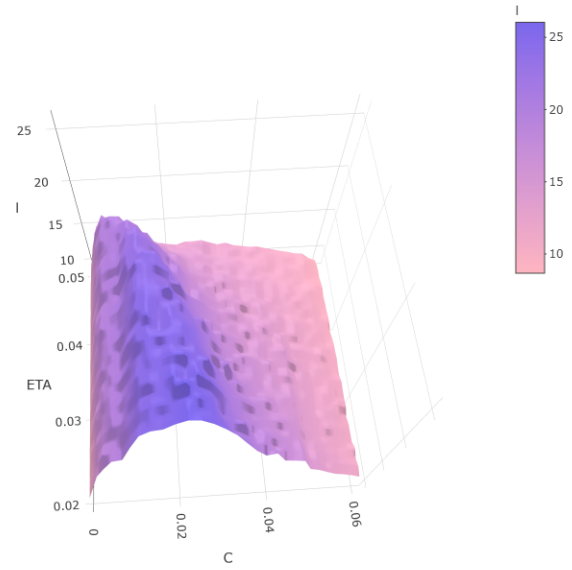

Figure 3.5: Simulation results of the number of significant groups when the group sizes are unbalanced

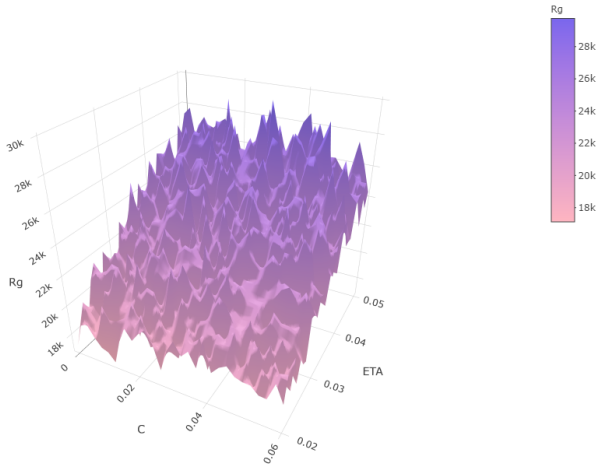

Figure 3.6: Simulation results of the total number of significant hypotheses when the group sizes are unbalanced

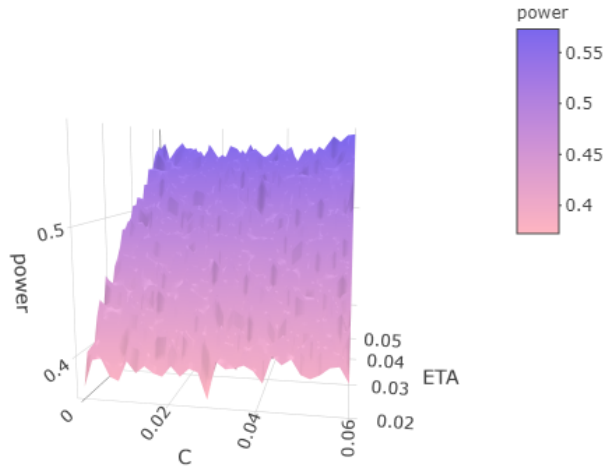

Figure 3.7: Simulation results of power when the group sizes are unbalanced

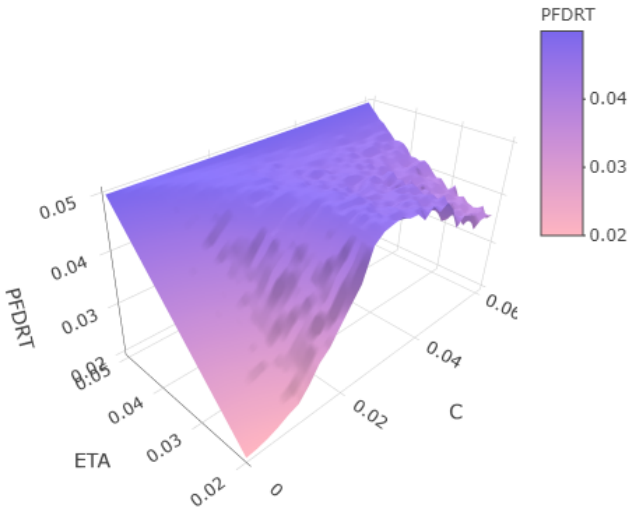

Figure 3.8: Simulation results of controlling the FDR when the group sizes are unbalanced

Chapter 4: REAL DATA APPLICATION 1

In this chapter, we apply the proposed method to groups of single nucleotide polymorphisms (SNPs). Bulik-Sullivan et al. (2015) and Finucane et al. (2015) partitioned the SNPs into groups with genomic features such as promoters and enhancers, and we select significant functional groups and SNPs within the selected groups. We select significantly associated pairs of the summary statistics from genome-wide association study (GWAS) and linkage disequilibrium (LD) score. We also perform LD score regression (LDSC) analysis, and we compare our results with those of LDSC. We further investigate additional local associations within haplotype blocks while existing methods such as LDSC use the whole SNPs.

4.1 Introduction to Linkage Disequilibrium Score Regression

Due to DNA microarrays for genotyping, scientists require simultaneous hypothesis tests for multiple regions of a genome in the twenty-first century. For example, using SNP, which is the simplest type of genetic marker, Speliotes et al. (2010) simultaneously tested the association between body mass index (BMI) and more than 1 million SNPs in GWAS. One of the new methods used to check the associations between phenotypes and the genetic effects is LDSC introduced by

Bulik-Sullivan et al. (2015). LD score is the numerical sum of the square of the pairwise correlations between SNP and neighboring SNPs, and LDSC focuses on the linear association between LD score and the summary statistics from GWAS. LDSC is known to distinguish confounding biases such as population stratification from heritability measuring the proportion of phenotypic variation due to the genetic differences. In particular, Finucane et al. (2015) extended LDSC to stratified LDSC to partition the heritability into groups. After partitioning the heritability by groups, Finucane et al. (2015) employed the multiple testing method to determine the significant groups which were enriched for heritability. However, the stratified LDSC does not explain which SNPs in the group cause the group to be discovered, so this dissertation is concerned with how to identify significant groups as well as SNPs within the groups. In particular, one major goal is to investigate the association between the summary statistics and LD score using the selected SNPs, whereas existing studies such as LDSC use the entire SNPs. Further, we examine the patterns of local associations within haplotype blocks. For this selection, we adopt the idea of combining two p-values computed from the summary statistic and LD score and then select significantly combined p-values based on multiple testing procedure controlling FDR. The direction we propose is different from the existing method in the following points. Our method reflects the idea of the sparsity of SNPs, so we exclude noise of SNPs and use only significant SNPs for regression analysis.

4.2 Sparse Signal Detection in Genetic Association Data

Under a polygenic model, LDSC is commonly used to estimate heritability and check genome-wide polygenic signals. Polygenicity means that the phenotype is affected by more than one functional SNPs with small effect sizes. Bulik-Sullivan et al. (2015) showed the polygenicity by observing the positive correlation between the summary statistics from GWAS and LD score using the whole SNPs. The result by the LDSC is made under the assumption that a large number of functional SNPs with small effect sizes provide genetic signals when they are aggregated in a biologically meaningful way. In such a case, we may not distinguish between hidden true genetic signals from the noise. In fact, LDSC does not distinguish the true and false null hypotheses since many of the SNPs are assumed to include signals although their strength may be fairly small. On the other hand, another common structural assumption in GWAS is that a majority of SNPs are simply noisy as mentioned in Wakefield (2008). Therefore, the LDSC approach may lead to some misleading results in case of a large number of noisy SNPs from the true null hypotheses. Unlike LDSC, under the assumption of sparsity of significant SNPs, we aim to discover genetic signals through explicit separation of true and false null hypotheses utilizing the proposed method discussed in the previous chapter. Specifically, we select significantly associated pairs of the summary statistics and LD score based on combining two p-values corresponding to the summary statistics and LD score. We apply regression analysis to the selected pairs of the summary statistic and LD score and additionally investigate the association within haplotype blocks. As a real

data application, we employ GWAS of BMI summary statistics from Speliotes et al. (2010) and examine the association between LD score and the summary statistics in the following sections.

4.3 Hypotheses for the Association Between GWAS Summary Statistics and LD score

According to the guidelines of Finucane et al. (2015) and Bulik-Sullivan et al. (2015), the SNPs of Speliotes et al. (2010) are classified into 52 groups, where SNPs can be included in more than one group. 644,055 distinct SNPs are used, and a total number of overlapped SNPs within 52 groups exceed 8 million. More specifically, there are m_g SNPs with a pair of GWAS summary statistics of BMI (s) and LD score (l) in the gth group for $1 \leq g \leq 52$ which are presented in Table 4.1.

We consider the jth SNP in the gth group provides (s_{gj}, l_{gj}) where s_{gj} measures the association between the corresponding SNP and BMI, and l_{gj} is the sum of the squared of the pairwise correlations between the corresponding SNP and other SNPs within the group.

Wakefield (2008) mentioned that the proportion of true non-null signals can be small in GWAS, known as sparse situation, which occurs commonly in large scale problems. We also consider the number of significant pairs of (s, l) is small in each group. Under the null hypothesis in the gth group (say, $H_{0,j}^{(g)}$), s_{gj} and l_{gj} are independent which are generated from probability density functions $f_g(s)$ and $h_g(l)$, respectively. Under the alternative hypothesis ($H_{1,j}^{(g)}$), s_{gj} and l_{gj} are generated from

g	m_g	ϵ_g	g	m_g	ϵ_g	g	m_g	ϵ_g	g	m_g	ϵ_g
1	24552	0.021	14	41094	0.015	27	203270	0.035	40	152894	0.006
2	80606	0.015	15	78952	0.036	28	38183	0.031	41	119873	0.044
3	34587	0.010	16	89102	0.021	29	121089	0.019	42	275924	0.024
4	304356	0.026	17	244734	0.031	30	201996	0.041	43	253251	0.037
5	19321	0.006	18	326365	0.030	31	300096	0.024	44	490819	0.029
6	53326	0.005	19	345769	0.034	32	314741	0.026	45	15210	0.018
7	131789	0.028	20	235433	0.024	33	7739	0.014	46	28429	0.029
8	424595	0.031	21	282240	0.028	34	27271	0.027	47	16599	0.023
9	115827	0.040	22	164541	0.028	35	30541	0.036	48	33030	0.008
10	170385	0.033	23	377376	0.025	36	36932	0.024	49	6796	0.019
11	403026	0.033	24	480855	0.026	37	247459	0.021	50	28701	0.032
12	4263	0.000	25	37941	0.040	38	403035	0.017	51	21014	0.007
13	16883	0.005	26	117990	0.034	39	150705	0.006	52	77139	0.028

Table 4.1: The detailed group information on the group sizes and the proportion of non-null SNPs for 52 groups

some joint probability distribution, denoted by $k_g(s,l)$. More specifically, for the gth group, the hypotheses are

$$H_{0,j}^{(g)} : (s,l) \sim f_g(s)h_g(l), \quad H_{1,j}^{(g)} : (s,l) \sim k_g(s,l).$$

As the marginal distributions, f_g and h_g under the null hypothesis $H_{0,j}^{(g)}$, we use

$$s_{gj} \sim \chi_1^2,$$

$$l_{gj} \sim lognormal(\mu_0^{(g)}, \sigma_0^{(g)}),$$

where χ_1^2 and $lognormal(\mu_0^{(g)}, \sigma_0^{(g)})$ means the chi-square distribution with degree of freedom 1 and lognormal distribution with parameters $\mu_0^{(g)}$ and $\sigma_0^{(g)}$. These assumptions of parametric distributions have been used in Finucane et al. (2015) and Cox et al. (2005), respectively.

Figure 4.1 shows the histograms of observed data and the null distributions of s_{gj} and l_{gj}, $1 \leq j \leq 304,356$ for the 4th group. In the case of chi-square and lognormal distributions, it is shown that they fit well with the actual data. This trend appears in all 52 groups, so it can be said that the chi-square distribution and lognormal distribution are appropriate to explain the distribution under the zero hypothesis of s and l.

In practice, the parameters $\mu_0^{(g)}$ and $\sigma_0^{(g)}$ in lognormal distribution for $l_{gj}, 1 \leq j \leq m_g$ are unknown, so we estimate $\mu_0^{(g)}$ and $\sigma_0^{(g)}$ using the data empirically which is the empirical null distribution used in Efron (2004). The estimation of the empirical null distribution is based on the zero assumption which implies that most of the data around the center are generated from the null distribution. Since only a small fraction of (s_{gj}, l_{gj}) is assumed to be significant, the number of significant l_{gj}s is also small among all l_{gj}s in each gth group. Furthermore, since l_{gj} is generated from lognormal distribution, the log-transformation of l_{gj}, $\log(l_{gj})$, follows the normal distribution with mean $\mu_0^{(g)}$ and standard deviation $\sigma_0^{(g)}$, so we can apply `locfdr`

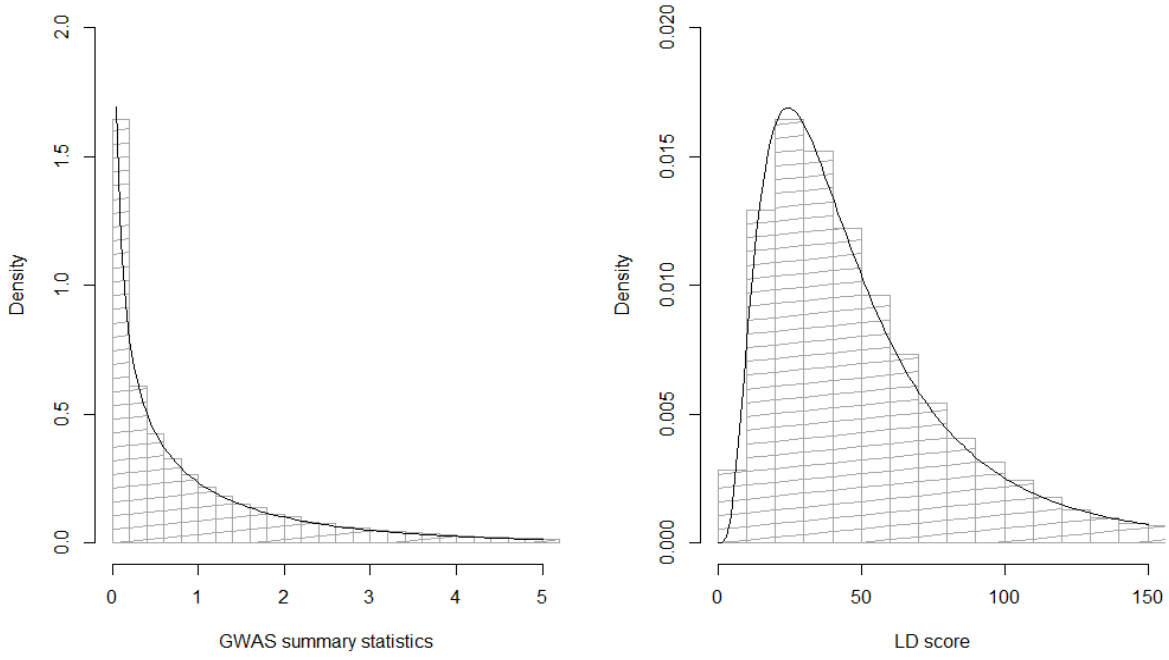

Figure 4.1: The histograms of observed data and the null distribution of GWAS summary statistics of BMI and LD score for the 4th group

procedure to $\log(l_{gj})$ and obtain estimators, $\hat{\mu}_0^{(g)}$ and $\hat{\sigma}_0^{(g)}$. From these marginal null distributions χ_1^2 and $lognormal(\hat{\mu}_0^{(g)}, \hat{\sigma}_0^{(g)})$, we can compute p-values which are

$$\begin{aligned} p_{sj}^{(g)} &= P(S > s_{gj}), \\ p_{lj}^{(g)} &= P(L_g > l_{gj}), \end{aligned}$$

where S and L_g are the random variables of χ_1^2 and $lognormal(\hat{\mu}_0^{(g)}, \hat{\sigma}_0^{(g)})$.

Since $p_{sj}^{(g)}$ and $p_{lj}^{(g)}$ are independent and each of them has uniform distribution on $(0, 1)$ under $H_{0,j}^{(g)}$, we consider combining these two p-values leading to a univariate p-value instead of bivariate p-values. There are several approaches to combining independent p-values such as Fisher's method in Fisher (1932) which combines $p_{sj}^{(g)}$

and $p_{lj}^{(g)}$ as follows:

$$\gamma_j^{(g)} = -2\log(p_{sj}^{(g)}) - 2\log(p_{lj}^{(g)}) \sim \chi_4^2, \tag{4.1}$$

where χ_4^2 is the chi-square distribution with degrees of freedom 4. The p-value corresponding to $\gamma_j^{(g)}$ is $p_j^{(g)} = P(\chi_4^2 > \gamma_j^{(g)})$ and by the probit transformation $x_{gj} = \Phi^{-1}(1 - p_j^{(g)})$, and thus we have independent and identically distributed standard normal distribution under $H_{0,j}^{(g)}$. Eventually, for the gth group, we have m_g test statistics, $(x_{g1}, x_{g2}, \cdots, x_{gm_g})$, $1 \leq g \leq G$ and the corresponding null hypotheses denoted by $(H_{0,1}^{(g)}, H_{0,2}^{(g)}, \cdots, H_{0,m_g}^{(g)})$.

4.4 The Results of the Proposed Multiple Testing Procedure

Similar to the simulation studies, we consider different values of the margin $M_g = cm_g, 1 \leq g \leq G$ and η. As we have done for the choices of c and η in Chapter 3 of simulation studies, we take the grid points of η from 0.02 to 0.05 and c from 0 to 0.06 to investigate the results corresponding to those different values of c and η. We apply the proposed method to test statistics obtained with (4.1) and examine the association between GWAS summary statistics of BMI and LD score. The results of Algorithm 2 are shown in Figure 4.2 - 4.4 analogous to the simulation studies with $\alpha = 0.05$. When c is relatively small such as $c \in [0, 0.01)$, almost all groups are selected as the significant groups regardless of η. With a larger value of c, we put a stronger requirement to declare significant groups, but the number of those significant groups tends to be diminished. From the definition of margin, it is expected that selected groups include a larger number of significant SNPs. Here, we

discover significant groups more selectively with large values of c greater than 0.01.

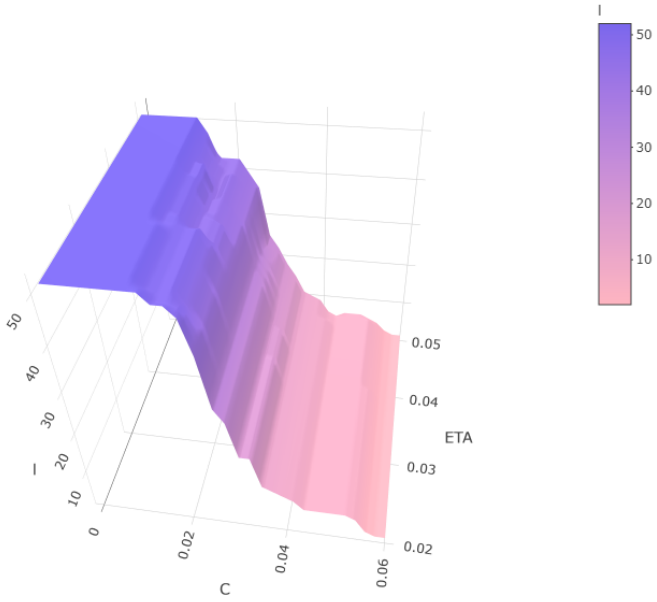

Figure 4.2: The results of the number of significant groups of Algorithm 2 to the test statistics combining GWAS summary statistics of BMI and LD score data

As mentioned in Remark 1, there may be a trade-off relationship between the number of significant groups and hypotheses within the group. To demonstrate this, we present Figure 4.5 for $c = 0.0331$ which shows that as η increases, the number of significant groups decreases, meanwhile the number of discovered hypotheses increases. In fact, as in this example, it is not necessarily the case of $\eta = \alpha$ that we obtain the maximum number of significant SNPs.

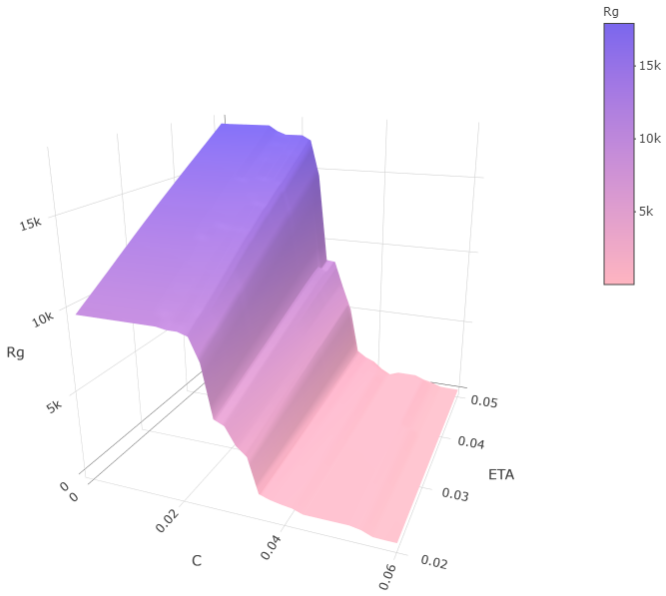

Figure 4.3: The results of the total number of rejected hypotheses of Algorithm 2 to the test statistics combining GWAS summary statistics of BMI and LD score data

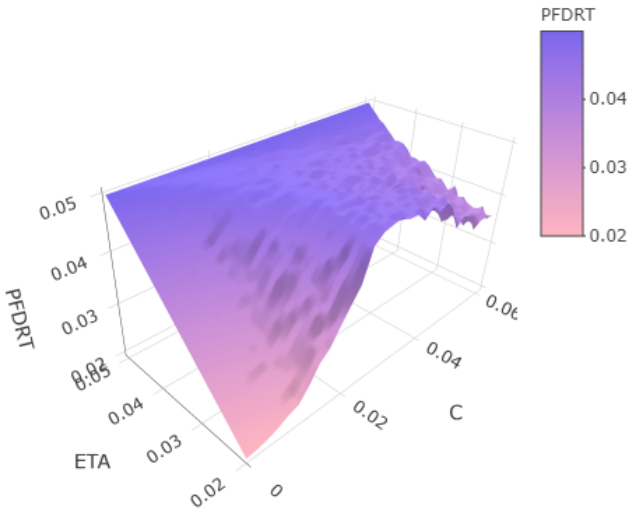

Figure 4.4: The results of controlling the FDR of Algorithm 2 to the test statistics combining GWAS summary statistics of BMI and LD score data

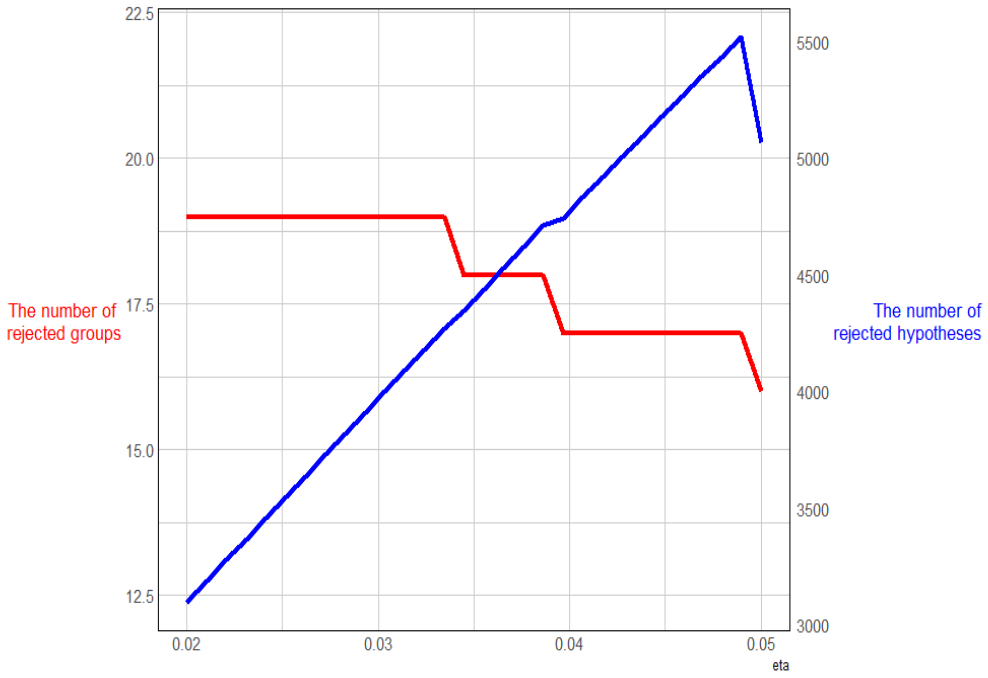

Figure 4.5: The trade-off relationship between the number of significant groups and hypotheses of the test statistics combining GWAS summary statistics of BMI and LD score data

4.5 LD Score Regression

When we select significant groups and SNPs within those groups for given c and η, we investigate the association between the summary statistics and LD score only for selected pairs. Many interesting relationships between GWAS summary statistics and LD score have been discovered by LDSC. Bulik-Sullivan et al. (2015) reported the polygenic effects that contributed to the positive slope of LDSC. On the other hand, Gazal et al. (2017) stated the summary statistics from GWAS were negatively correlated with LD score under specific model assumptions. In our case, we perform regression analysis with selected pairs of summary statistics and LD score, and then the results are compared with the result of LDSC in Bulik-Sullivan et al. (2015) which is based on all pairs of summary statistics and LD score.

We present Figure 4.6 and Figure 4.7 to compare two approaches such as regression analysis for the 4th group with selected or all pairs of the summary statistics and LD score. Figure 4.6 represents the results of LDSC with the SNPs within the 4th group following the method in Bulik-Sullivan et al. (2015). Each point in Figure 4.6 represents an LD score quantile where the x coordinate of the point is the weighted mean of LD score using the SNPs in that quantile, and the y coordinate of the point is the weighted mean of the summary statistics employing the SNPs in that quantile. The black solid line in Figure 4.6 is the fitted regression line. The genomic control inflation factor is 1.1175, and the intercept of LDSC is 0.788.

In contrast, Figure 4.7 shows the discovered SNPs (+) as the result of Algo-

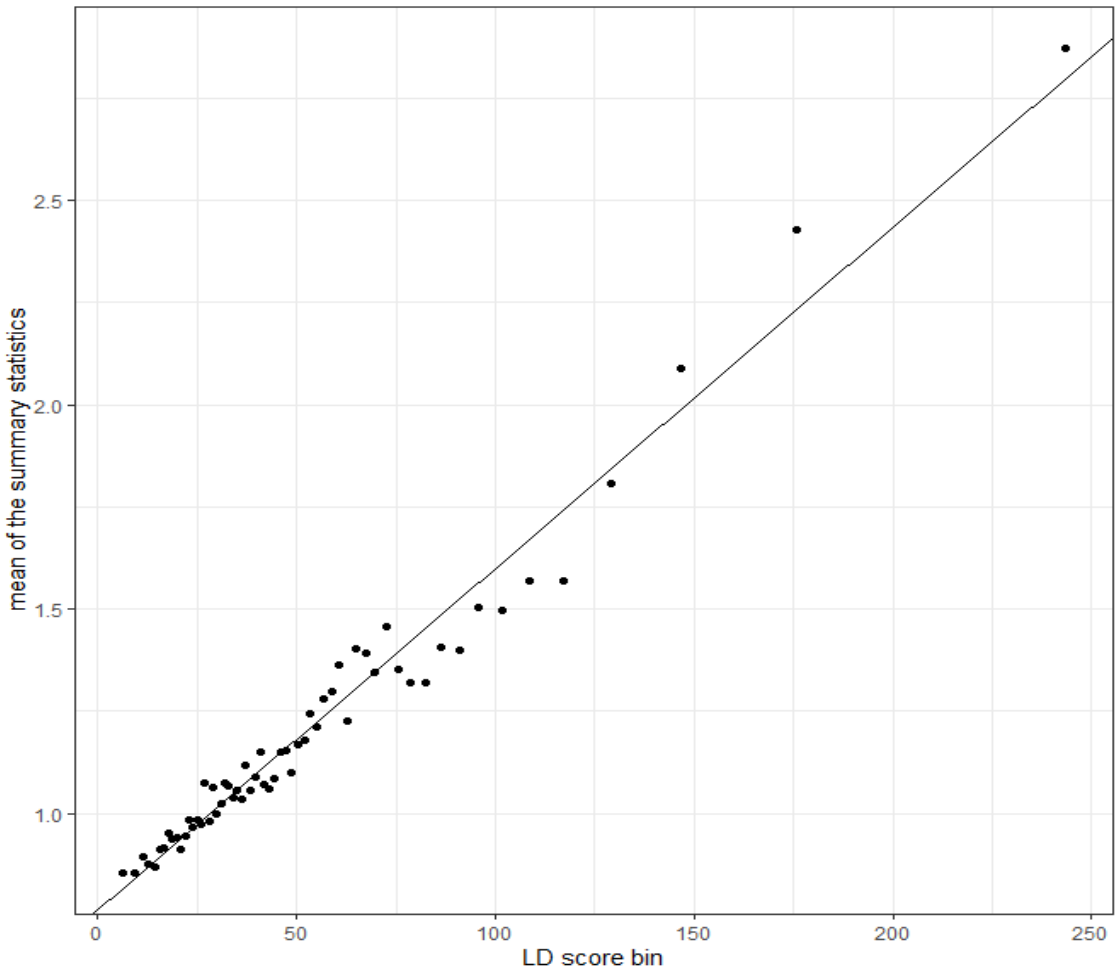

Figure 4.6: The result of LDSC for all SNPs in the 4th group

rithm 2 with $c = 0$ and $\eta = 0.05$ in the 4th group. The solid line in Figure 4.7 is the regression line estimated with all SNPs, and the dashed line in Figure 4.7 is the fitted regression line using only the significant SNPs. By considering Figure 4.7 and Figure 4.6, it can be seen that the signs of the regression lines obtained using all SNPs and that of the discovered SNPs are opposite. In other words, LDSC hypothesized that a large number of functional SNPs with small quantitative effect sizes affect BMI through significant associations between LD score and the summary statistics. However, the associations observed by the statistically significant pairs

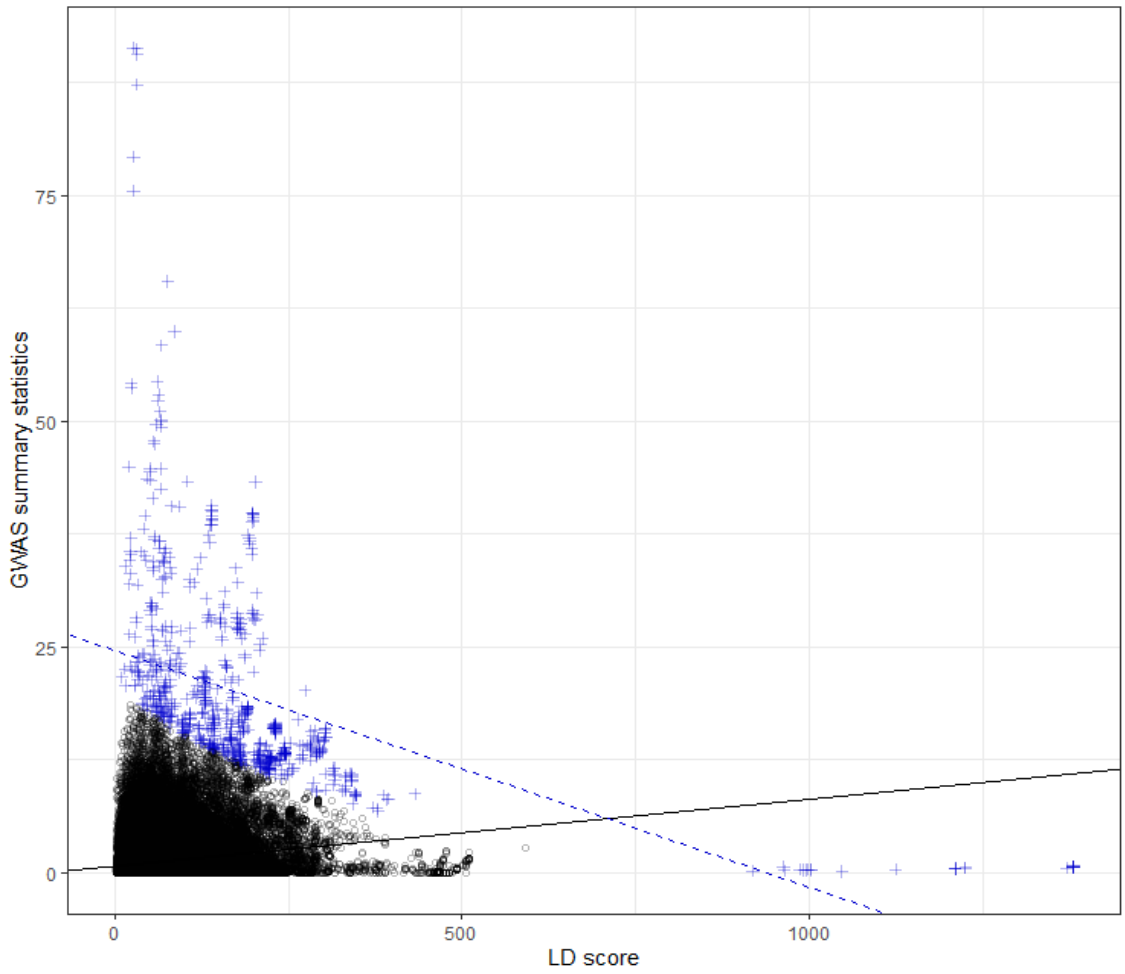

Figure 4.7: Pairs of LD score and the summary statistics for the selected SNPs with fitted regression lines

of LD score and the summary statistics are different from the results of LDSC. In Figure 4.7, as LD score increases, the summary statistics tend to decrease due to their negative association. Specifically, large values of summary statistics with small values of LD score represent the SNPs which affect BMI independently, while SNPs with relatively small summary statistics are correlated with other SNPs since the LD scores increase. This motivates further how statistically significant SNPs can be used to obtain the same pattern of association with LDSC.

4.6 The Local Association Between GWAS Summary Statistics and LD score

In the previous section, we discuss the association between the summary statistic and LD score using the selected SNPs by the proposed multiple testing. The discovered SNPs in the 4th group are distributed across the genome, and additional information on genomic location of SNPs is available to gain insights into selective local association between the summary statistics and LD score. Specifically, since SNPs are highly associated within local regions within chromosome, called haplotype blocks which was noted by Slatkin (2008), it is of interest to investigate and analyze the association between LD score and the summary statistics of BMI per unit of haplotype block. To characterize local association between LD score and the summary statistics, we consider the haplotype blocks of the selected SNPs, and we explore genetic dependence by narrowing the scope of the intra-chromosome.

Figure 4.8-4.10 show three different types of association patterns in regression analysis for haplotype blocks from different chromosomes in the 4th group as a result of Algorithm 2 with $c = 0$ and $\eta = 0.05$.

Figure 4.8 shows the regression results of the selected SNPs providing pairs of the summary statistics and LD score on chromosome 12 in the 4th group. The right panel in Figure 4.8 represents the structure of three haplotype blocks, and the left panel of Figure 4.8 shows three dashed regression lines to illustrate the local association between the summary statistics and LD score, for each of the haplotype blocks.

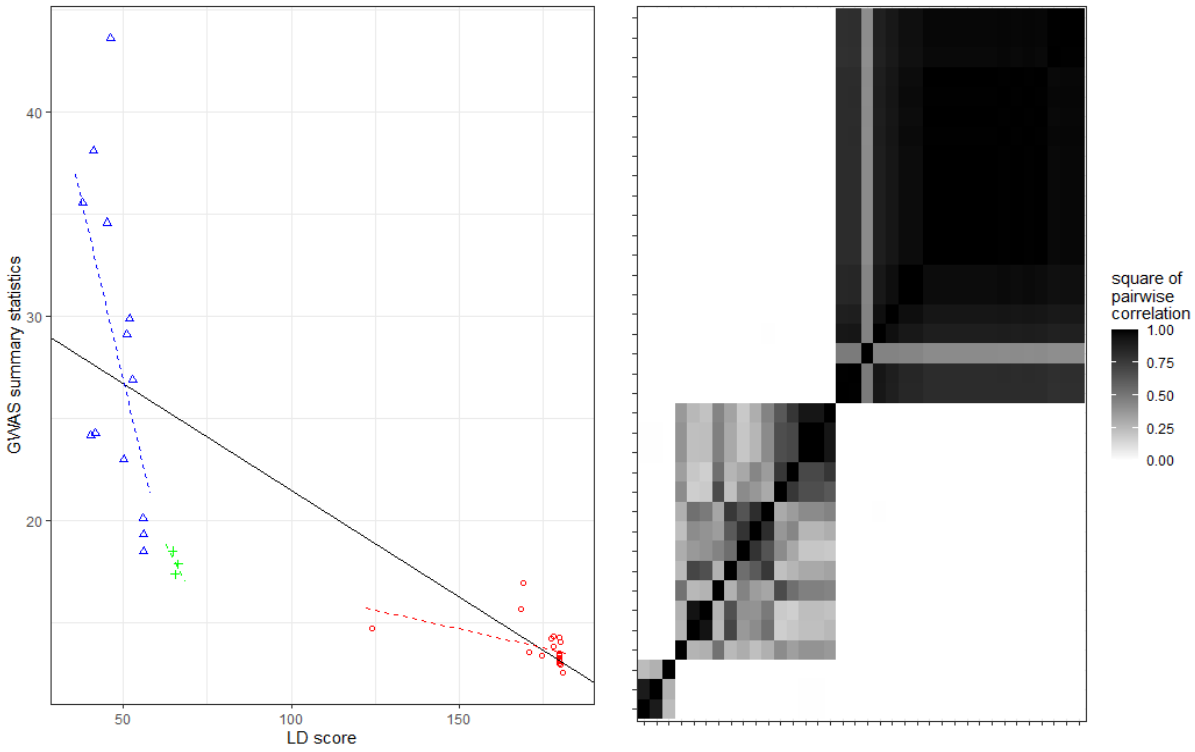

Figure 4.8: The result of the selected SNPs providing pairs of the summary statistics and LD score on chromosome 12 in the 4th group

The solid line in Figure 4.8 means the fitted regression line using all selected SNPs on chromosome 12 in the 4th group, and all the slopes of the regression analyses support a negative correlation between LD score and the summary statistics. This example shows that the associations within haplotype blocks have the same trend as the association of all selected pairs of the summary statistics and LD score.

Figure 4.9 represents the regression results of the selected SNPs providing pairs of the summary statistics and LD score on chromosome 15 in the 4th group in the same way as Figure 4.8. In this case, we have two haplotype blocks in which we have different signs of slopes in two regression lines. Moreover, when we ignore the haplotype block structure and use all selected SNPs on chromosome 15 in the 4th

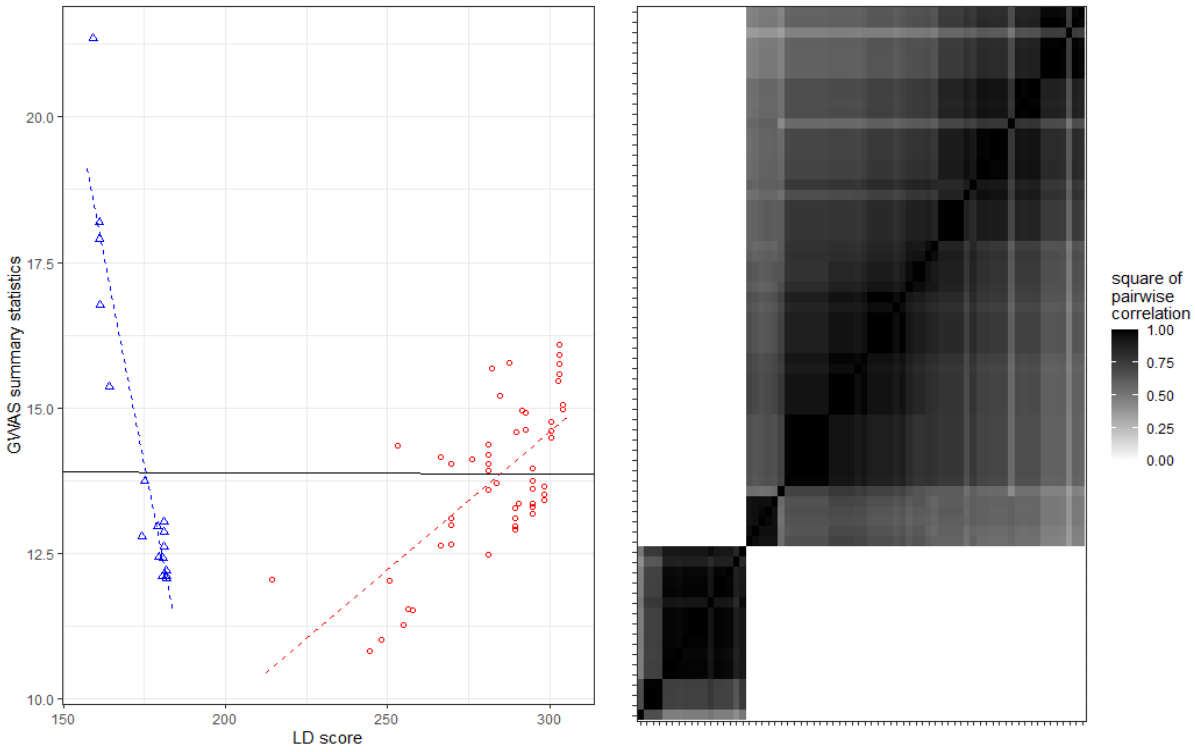

Figure 4.9: The result of the selected SNPs providing pairs of the summary statistics and LD score on chromosome 15 in the 4th group

group to estimate the regression line, the linear relationship between the LD score and the summary statistics almost disappears. From this example, we see that there may exist opposite signs of slopes in regression lines in haplotype blocks and such different signs of associations may wash out overall association.

Another pattern of local association is shown in Figure 4.10, which is the case of chromosome 11 in the 4th group. Similar to Figure 4.8, we observe four haplotype blocks in the right panel of Figure 4.10. Here, LD scores and the summary statistics are positively correlated for the selected SNPs in the three haplotype blocks except one case of one block with a negligible association. On the other hand, the regression line using all selected SNPs on chromosome 11 in the 4th group shows a negative

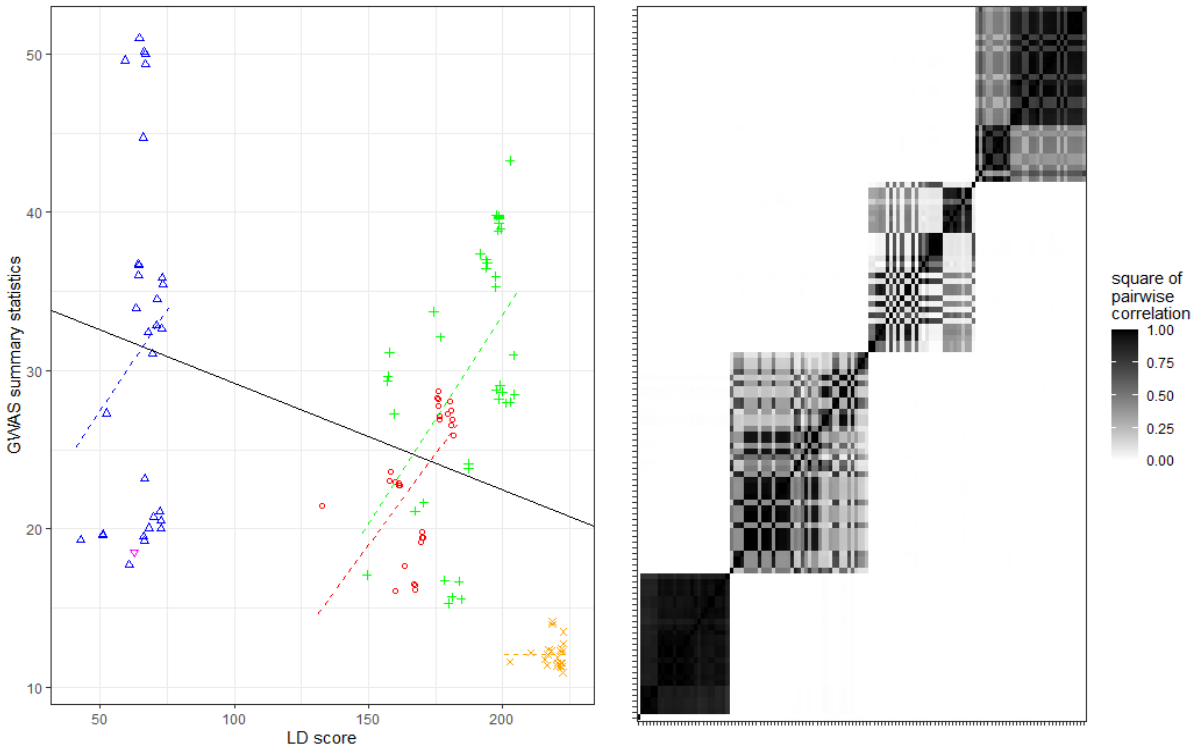

Figure 4.10: The result of the selected SNPs providing pairs of the summary statistics and LD score on chromosome 11 in the 4th group

correlation between LD score and the summary statistics. This example is the case that the association of overall selected pairs have different sign of slope than the slopes in regression lines from haplotype blocks.

We observe the various patterns of selective local associations between LD score and the summary statistics depending on the size and location of haplotype blocks. While the choice of the haplotype blocks for the local association may deserve further study, it is clear from these applications that by aggregating and using the entire SNPs, various patterns of the local association in intra-chromosomes between LD score and the summary statistics disappear. This phenomenon can be explained by Simpson's paradox mentioned in Blyth (1972), in which a linear trend in a specific

direction appears in data of several groups, but disappears or reverses when the data are combined.

It is worth noting that our method and Finucane et al. (2015) are based on partitioning SNPs into groups, so the inferential results can be considerably affected by how the groups are formed. Specifically, in the right panel in Figure 4.8, we see a set of the selected SNPs with strong squared pairwise correlations within the haplotype block. The inclusion of many SNPs from haplotype block with strong squared pairwise correlations in the data can lead to confounding biases in discovering true genetic signals.

Chapter 5: REAL DATA APPLICATION 2

In this chapter, we apply the proposed method to groups of SNPs in deoxyribonuclease I (DNase I) hypersensitive sites (DHSs) to see the association with type 2 diabetes. A number of studies have been conducted for the association between human organs and type 2 diabetes, particularly the pancreas. Here, among the 32 groups of DHSs including a group related to pancreas, we select significantly associated DHSs groups to type 2 diabetes. When the margin is small, all DHSs groups are determined to be a significant group, and it is hard to identify the groups truly associated with type 2 diabetes. On the other hand, when the margin is sufficiently large, the proposed method selects the DHSs group related to pancreas as the most significantly associated group to type 2 diabetes. Four of DHSs groups are finally selected with conservative group selection criteria, and the results are consistent with existing medical literature.

5.1 Introduction to Encyclopedia of DNA Elements Projects

GWAS has been used as a useful tool to identify common genetic variants associated with complex traits or diseases. In particular, SNPs are the most common type of genetic variation used in GWAS, and each SNP represents a single base-pair

difference in the DNA sequence. Most of the SNPs significantly associated with traits or diseases identified by GWAS are within the non-coding region, and most of these non-coding variants are concentrated in DHSs (Giral et al., 2018). DHSs are specific regions with increased chromatin accessibility, and DHSs are known to be a group of generic markers of regulatory DNA. Thus, DHSs can be used to build a better understanding of gene regulatory networks, the organization and functions of the human.

The Encyclopedia of DNA Elements (ENCODE) Project is a research project that seeks to interpret the human or mouse genome sequence, and aims to identify functional elements of the human or mouse genome. For example, phase three of the ENCODE Project performed analysis of the cell and tissue repertoires of RNA transcription, chromatin structure and modification, DNA methylation, chromatin looping, and occupancy by transcription factors and RNA-binding proteins (Moore et al., 2020). Davis et al. (2018) provided the ENCODE Portal which is a freely accessible database and website as a source of data generated by the ENCODE Consortium. ENCODE Portal has provided a large number of sequencing libraries from assays including RNA-Seq and DNase-Seq. In the ENCODE Portal, we are able to find various DHSs in the human genome. Furthermore, Meuleman et al. (2020) produced a comprehensive collection of high-resolution maps of DHSs obtained from 733 human bio-samples within the human genome sequence including regions related to the major organ systems. Among 733 DHSs, we choose 32 DHSs related to distinct human organs to identify significant groups associated with type 2 diabetes.

GWAS in samples of European ancestry performed simultaneous association

tests between more than nine millions of SNPs and type 2 diabetes in Pan-ancestry genetic analysis of the UKB (2020). Using a generalized mixed model association testing framework, each SNP had a regression coefficient and corresponding p-value and genetic location. Based on the genetic locations of SNPs used in GWAS, we grouped the SNPs into selected 32 DHSs using `GenomicRanges` in R-package. Group information and downloaded 32 data sets of DHSs from the ENCODE Portal (Davis et al., 2018) can be found in Table 5.1 - 5.2.

	Organ	Description	ENCODE accession numbers	the number of SNPs
1	Adrenal Gland	Adrenal	ENCFF525FRH	57749
2	Amion	Amniotic Epithelial	ENCFF316LBU	166256
3	Blood	CD34+ Hematopoietic progenitor cells	ENCFF059VUS	159512
4	Bone	Bone leg right	ENCFF369GLM	119537
5	Brain	Brain	ENCFF528GDM	152135
6	Esophagus	Esophageal Epithelial	ENCFF527EJQ	174233
7	Eye	Choroid Plexus Epithelial	ENCFF719FXT	169100
8	Gonad	Testes	ENCFF843ZSC	52395
9	Gum	Gum fibroblast	ENCFF507FBF	39447
10	Heart	Cardiomyocytes	ENCFF156RTG	151188
11	Kidney	Kidney	ENCFF403LQM	103415
12	Large Intestine	Intestine Lg	ENCFF920DDN	82457
13	Liver	h.f.liver	ENCFF286LYP	62734
14	Lung	Lung Fibroblasts	ENCFF331SYD	173580
15	Mammary	Mammary Fibroblasts	ENCFF387KMX	58274
16	Mesoderm	CD3	ENCFF216NXR	26491

Table 5.1: The detailed group information and downloaded 16 data set of DHSs from the ENCODE Portal

	Organ	Description	ENCODE accession numbers	the number of SNPs
17	Muscle	Muscle leg	ENCFF674ZTX	193786
18	Ovary	Ovary	ENCFF883WWT	39934
19	Pancreas	Pancreas	ENCFF897PRD	36541
20	Periodontal Ligament	Periodontal Ligament Fibroblasts	ENCFF620HKT	139303
21	Placenta	Villous Mesenchymal Fibroblasts	ENCFF311UAO	145779
22	Prostate	Human Prostate Epithelial Cells	ENCFF608WCU	48507
23	Pulmonary Artery	Pulmonary Artery Fibroblasts	ENCFF385ZNB	134827
24	Skin	Dermal Fibroblasts	ENCFF445GCV	169356
25	Small Intestine	Intestine Sm	ENCFF442AYJ	33700
26	Spinal Cord	Human Astrocytes - spinal cord	ENCFF674LBB	132947
27	Spleen	Spleen	ENCFF587YNA	31294
28	Stomach	Stomach	ENCFF933ABR	185153
29	Stroma	Bone marrow stromal cells	ENCFF254WCU	75282
30	Tongue	Tongue	ENCFF173CEG	157060
31	Umbilical	Umbilical vein endothelial	ENCFF097KBE	101275
32	Vascular	Human Brain Vascular Smooth Muscle Cells	ENCFF788MXD	31566

Table 5.2: The detailed group information and downloaded 16 data set of DHSs from the ENCODE Portal

5.2 Enrichment Test

Maurano et al. (2012) performed simultaneous association tests between DHSs groups of SNPs and diseases such as Crohn's disease, multiple sclerosis, and QRS duration. Considering the enrichment of DHSs groups, Maurano et al. (2012) determined strongly associated DHSs groups with these diseases. The enrichment of

DHSs groups is defined as the ratio of the proportion of significant SNPs in the DHSs group to the proportion of significant SNPs in data. As changing p-value thresholds, Maurano et al. (2012) observed the patterns of the enrichment of the DHSs groups. However, Maurano et al. (2012) did not consider the multiplicity correction to the calculation of the enrichment of DHSs groups. In this dissertation, instead of measuring the enrichment without considering the multiplicity problem, we apply the proposed method to selected 32 groups of DHSs, and we measure the relative importance of groups utilizing various group significance thresholds illustrated by the margin.

5.3 The Results of the Proposed Multiple Testing Procedure

In this section, we apply the proposed method to the data obtained from ENCODE Portal and GWAS summary statistics. Each SNP has a genomic location and provides p-value which is the measurement of the linear association between the corresponding genetic variation and type 2 diabetes. Using the genomic locations, the SNPs are partitioned into the selected 32 DHSs groups. With the probit transformation utilized in Chapter 4, we obtain a test statistic for each SNP for type 2 diabetes. Here our goal is to detect the significantly related DHSs groups represented by the human organ system for type 2 diabetes.

Type 2 diabetes is the most common type of diabetes and is a chronic disease with high blood glucose levels. The pancreas is the organ which produces insulin, one of the main hormones that helps to regulate blood glucose levels (Marchetti

et al., 2017a). If someone has type 2 diabetes, insulin resistance prevents other organs, such as the liver and muscle cells, from responding properly to insulin (Taylor, 2012). The pancreas produces more insulin to control the body, however it cannot meet the increased demand sometimes. When the pancreas can no longer produce sufficient insulin for lower blood glucose levels, symptoms of diabetes begin to appear. Therefore, it can be said that the pancreas is the most associated organ for type 2 diabetes.

The proposed algorithm requires two given thresholds. In particular, the two given thresholds are η, which controls the local FDR within the group, and the margin that determines the condition for being an important group in a group significance test. Here η is fixed at $\alpha = 0.05$, and the margin is set proportional to the number of SNPs in the group, i.e., $M_g = cm_g, 1 \leq g \leq 32$ where m_g is the number of SNPs in the gth DHSs group and c is a given constant. In other words, to be a significant group, the larger the group size, the larger the number of significant SNPs is required. The significance of the DHSs groups are evaluated while changing c, which controls the margin, we set c from 0 to 0.0525. Figure 5.1 shows the results of the number of significant groups selected by the proposed method with controlling for FDR at level α. As c increases, the number of selected groups decreases due to the stronger requirement for the genetic signal within the selected DHSs group. Figure 5.2 represents the rank of relative importance to the selected DHSs groups as the margin increases.

In the Figure 5.2, the x-axis represents c which controls the margin, and the y-axis represents the sorted order of DHSs groups in ascending order from 1 to 32. We

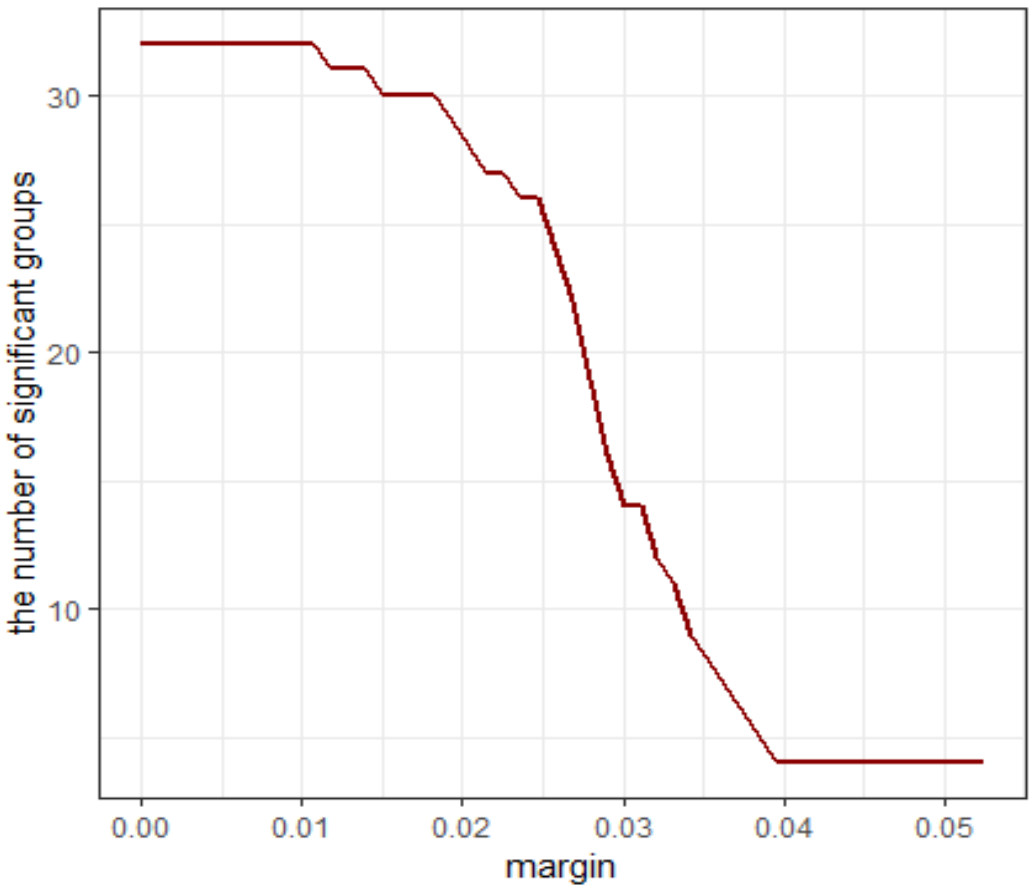

Figure 5.1: The number of selected DHSs groups according to the margin

can see the flow line tracing the path of the group significance as the margin increases in the Figure 5.2. If a DHSs group is not selected as a significant group, the path is discontinuous. When c is smaller than 0.01, so the margin is small, all 32 DHSs groups are selected as significant groups. In the group selection point of view, the results of 32 significant DHSs groups are not informative. As the margin increases, the number of significant groups decreases. When c is greater than 0.04, only four DHSs groups are selected as important groups, and their order of importance is unchanged. In particular, the DHSs group for pancreas is chosen by the method as the 6th most important group for type 2 diabetes when the margin is small or

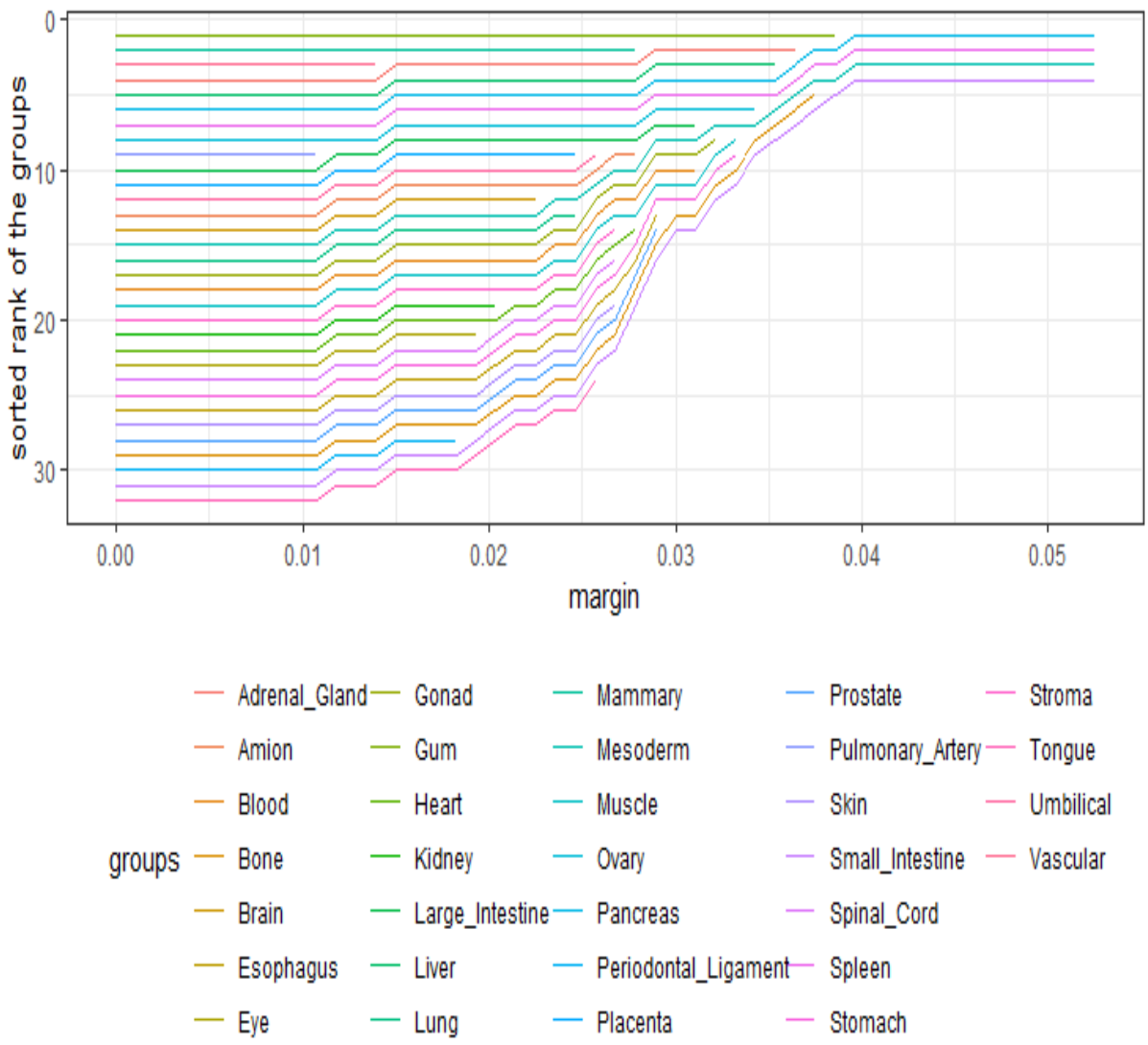

Figure 5.2: The rank of the selected 32 DHSs groups according to the margin

c is less than 0.015. The group is selected as the most important group when the margin is large or c is greater than 0.0396.

When the margin is large or c is more than 0.0396, the selected DHSs groups are related to pancreas, spleen, mesoderm, and small intestines. Many researches have been conducted for the association between human organs and type 2 diabetes. For example, the association between pancreas and type 2 diabetes is illustrated in

Marchetti et al. (2017b), and the association between small intestines is described in Sanyal (2013). The SNPs in mesoderm are related to CD3 primary Cells and CD3 protein complex is an important T cell marker for the immune system. The spleen is a part of the lymphatic system and stores as well as filters blood and makes white blood cells that protect the body from infection. De Candia et al. (2019) and Berbudi et al. (2020) mentioned the association between immune system and type 2 diabetes.

Chapter 6: SUMMARY AND FUTURE WORK

We have discussed the selection of significant groups and hypotheses within such significant groups using multiple testing procedures controlling FDR. The proposed multiple testing procedure is based on introducing the margin for group significance leading to more selectively chosen groups and hypotheses within those selected groups. Many statistical hypothesis testing methods using the idea of margin have been developed in many areas, however to the best of our knowledge, we first introduce this idea to group sequential multiple testing problems.

As a real data application, we present the results of regression analysis utilizing GWAS of BMI data within haplotype blocks based on the selected SNPs, and a few ideas on our proposed procedure. To observe associations between the summary statistics and LD score, we use the combined p-values method and then select significantly associated pairs of the summary statistics and LD score. One conventional approach such as LDSC considers the association of all pairs of summary statistics and LD scores, since it is based on the idea that aggregation of all SNPs increases the effect of SNPs on phenotype. This is in contradiction to some other studies such as Wakefield (2008) claiming that only a fraction of SNPs affects the phenotype. In this sense, the proposed regression using selected pairs follows the

idea of structural assumption such as sparsity of effect of SNPs on phenotype. By attempting regression analysis only on selected pairs, it is shown that a new pattern such that LD score and the summary statistic tend to have negative correlations. Further, through regression within the haplotype block, we have demonstrated that the summary statistic and LD score have more diverse patterns. Since haplotype blocks may correspond to hidden structures, regression analysis without such sub-structures may have misleading results such as Simpson's paradox. Therefore, we claim that it is meaningful to investigate local association of the summary statistics and LD score which may shed a new light on the area of LDSC.

In the second real data application, among selected 32 DHSs groups, we detect the statistically associated DHSs groups with type 2 diabetes utilizing the proposed method. Intuitively, the pancreas, the organ that produces insulin, has the most significant association with type 2 diabetes. When the number of DHSs groups is large, it is difficult to selectively choose a significant group with a small margin, however when the margin is large, the proposed model selects the pancreas as the most important group. The results mean that the global test for testing the complete null hypothesis is not effective when the number of groups is large as well as the number of hypotheses within the group is large. Through literature on the four selected DHSs groups and organs under the conservative group selection thresholds, we confirm the validity of the proposed methodology.

The proposed method ensures that each group has its own proportion of alternative hypotheses and empirical null distribution. However, our main interest is the characteristics of the selected groups. For future work, common proportion

parameter and null distribution to non-significant groups would be a extension of our current work. EM algorithm (Dempster et al., 1977) can be one approach to estimate the common features. Furthermore, we consider other highly heritable traits or diseases. Specifically, test statistics are affected by disease, and LD scores are dependent on population. Discovering SNPs for various diseases along with LD score helps to understand the association between genetic dependent architecture within a population and disease. Finally, we are able to perform enrichment tests. When η increases, we can observe the pattern of concentration of significant SNPs in each group and in data. By doing so, we are able to improve the enrichment test used in (Maurano et al., 2012) more rigorously.

Appendix A: SUPPLEMENTARY MATERIALS

The supplementary material provides some technical details.

A.1 Derivation of Likelihood Function in (3.6):

To proceed further with (3.6), consider $\Omega_g = \{0,1\}^{m_g}$ which is a sample space of $\boldsymbol{\theta}_{j|g}$, $1 \leq g \leq G, 1 \leq j \leq m_g$. Ω_g is partitioned into two disjoint sets with given M_g: Ω_{g0}, Ω_{g1}. If $\theta_g = 0$ is a gth group condition, then $\boldsymbol{\theta}_{j|g} \in \Omega_{g0}$ where $\Omega_{g0} = \{\boldsymbol{\theta}_{j|g} : \sum_{j=1}^{m_g} \theta_{j|g} \leq M_g\}$. On the other hand, if $\theta_g = 1$ is a gth group condition, then $\boldsymbol{\theta}_{j|g} \in \Omega_{g1}$ where $\Omega_{g1} = \{\boldsymbol{\theta}_{j|g} : \sum_{j=1}^{m_g} \theta_{j|g} > M_g\}$. These sample spaces have the same meaning as the indicator functions in (3.2)-(3.3). Furthermore, if the gth group and the jth hypothesis within the group are significant, an additional sample space $\Omega^*_{g1} = \{\boldsymbol{\theta}_{k|g} : \sum_{k,k\neq j}^{m_g} \theta_{k|g} > M_g - 1\}$ is also considered. Define

$$f_{\theta_{j|g}}(x_{gj}) = (1 - \theta_{j|g}) f_{0,g}(x_{gj}) + \theta_{j|g} f_{1,g}(x_{gj}),$$

$$f_g(x_{gj}) = (1 - \epsilon_g) f_{0,g}(x_{gj}) + \epsilon_g f_{1,g}(x_{gj}),$$

where $f_{\theta_{j|g}}$ is the conditional density of x_{gj} for a given latent variable $\theta_{j|g} = 0$ or 1 and f_g is the marginal density of x_{gj}. The conditional probability of $\mathbf{x}_g$ given $\theta_g = 1$

is

$$
\begin{aligned}
&P(\mathbf{x}_g|\,\theta_g = 1) \\
&= \sum_{\boldsymbol{\theta}_{j|g}\in\Omega_{g1}} \left(\prod_{j=1}^{m_g} \frac{\epsilon_g^{\theta_{j|g}}(1-\epsilon_g)^{1-\theta_{j|g}}}{P(\sum_{k=1}^{m_g}\theta_{k|g} > M_g)} f_{\theta_{j|g}}(x_{gj}) \right) \\
&= \frac{1}{P(\sum_{k=1}^{m_g}\theta_{k|g} > M_g)} \sum_{\boldsymbol{\theta}_{j|g}\in\Omega_{g1}} \left(\prod_{j=1}^{m_g} \epsilon_g^{\theta_{j|g}}(1-\epsilon_g)^{1-\theta_{j|g}} \times \{(1-\theta_{j|g})f_{0,g}(x_{gj}) + \theta_{j|g} f_{1,g}(x_{gj})\} \right) \\
&= \frac{1}{P(\sum_{k=1}^{m_g}\theta_{k|g} > M_g)} \sum_{\boldsymbol{\theta}_{j|g}\in\Omega_{g1}} \left(\prod_{j=1}^{m_g} \widetilde{fdr}_{gj}^{\,1-\theta_{j|g}}(1-\widetilde{fdr}_{gj})^{\theta_{j|g}} \right) \prod_{j=1}^{m_g} f_g(x_{gj}) \\
&\approx \frac{\prod_{j=1}^{m_g} f_g(x_{gj})}{P(\sum_{k=1}^{m_g}\theta_{k|g} > M_g)}\, P(\sum_{j=1}^{m_g}\theta_{j|g} > M_g|\,\mathbf{x}_g). \qquad \text{(A.1)}
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&P(\mathbf{x}_g|\,\theta_g = 0) \\
&= \sum_{\boldsymbol{\theta}_{j|g}\in\Omega_{g0}} \left(\prod_{j=1}^{m_g} \frac{\epsilon_g^{\theta_{j|g}}(1-\epsilon_g)^{1-\theta_{j|g}}}{P(\sum_{k=1}^{m_g}\theta_{k|g} \le M_g)} f_{\theta_{j|g}}(x_{gj}) \right) \\
&= \frac{1}{P(\sum_{k=1}^{m_g}\theta_{k|g} \le M_g)} \sum_{\boldsymbol{\theta}_{j|g}\in\Omega_{g0}} \left(\prod_{j=1}^{m_g} \epsilon_g^{\theta_{j|g}}(1-\epsilon_g)^{1-\theta_{j|g}} \times \{(1-\theta_{j|g})f_{0,g}(x_{gj}) + \theta_{j|g} f_{1,g}(x_{gj})\} \right) \\
&= \frac{1}{P(\sum_{k=1}^{m_g}\theta_{k|g} \le M_g)} \sum_{\boldsymbol{\theta}_{j|g}\in\Omega_{g0}} \left(\prod_{j=1}^{m_g} \widetilde{fdr}_{gj}^{\,1-\theta_{j|g}}(1-\widetilde{fdr}_{gj})^{\theta_{j|g}} \right) \prod_{j=1}^{m_g} f_g(x_{gj}) \\
&\approx \frac{\prod_{j=1}^{m_g} f_g(x_{gj})}{P(\sum_{k=1}^{m_g}\theta_{k|g} \le M_g)} P(\sum_{j=1}^{m_g}\theta_{j|g} \le M_g|\,\mathbf{x}_g). \qquad \text{(A.2)}
\end{aligned}
$$

The normal approximations in (A.1) and (A.2) are from (3.8) leading to the approximate likelihood function as follows:

$$
\begin{aligned}
&L(\boldsymbol{\Phi}|\mathbf{x},\boldsymbol{\Theta}) \\
&\approx \prod_{g=1}^{G} \left(\frac{\pi_1 P(\sum_{j=1}^{m_g}\theta_{j|g} > M_g|\,\mathbf{x}_g)\prod_{j=1}^{m_g} f_g(x_{gj})}{P(\sum_{k=1}^{m_g}\theta_{k|g} > M_g)} + \frac{(1-\pi_1)P(\sum_{j=1}^{m_g}\theta_{j|g} \le M_g|\,\mathbf{x}_g)\prod_{j=1}^{m_g} f_g(x_{gj})}{P(\sum_{k=1}^{m_g}\theta_{k|g} \le M_g)} \right).
\end{aligned}
$$

A.2 Derivation of Algorithm 1

Consider the data $\mathbf{x} = \{x_{gj}\}$ and latent variables $\mathbf{\Theta} = (\theta_g, \theta_{j|g})$ for all (g, j) as the complete data. The complete likelihood function of $\mathbf{\Phi} = (\pi_1, \epsilon_g)$ for EM algorithm can be written as

$$L(\mathbf{\Phi}|\mathbf{x}, \mathbf{\Theta}) = \prod_{g=1}^{G} \Bigg\{ \Big(\pi_1 P(\mathbf{x}_g|\theta_g = 1) \Big)^{I(\theta_g=1)} \Big((1-\pi_1) P(\mathbf{x}_g|\theta_g = 0) \Big)^{I(\theta_g=0)} \Bigg\},$$

leading to the log-likelihood function

$$\begin{aligned}
l(\mathbf{\Phi}|\mathbf{x}, \mathbf{\Theta}) &= \sum_{g=1}^{G} \Bigg\{ I(\theta_g = 1) log \Big(\pi_1 P(\mathbf{x}_g|\theta_g = 1) \Big) + I(\theta_g = 0) log \Big((1-\pi_1) P(\mathbf{x}_g|\theta_g = 0) \Big) \Bigg\} \\
&= \sum_{g=1}^{G} I(\theta_g = 1) log(\pi_1) + \sum_{g=1}^{G} I(\theta_g = 0) log(1-\pi_1) \\
&+ \sum_{g=1}^{G} \sum_{j=1}^{m_g} \Big(I(\theta_{j|g} = 1) log(\epsilon_g) + I(\theta_{j|g} = 0) log(1-\epsilon_g) \Big) \\
&+ \sum_{g=1}^{G} \sum_{j=1}^{m_g} \Bigg(I(\theta_g = 1, \theta_{j|g} = 1) log f_{1,g}(x_{gj}) + I(\theta_g = 1, \theta_{j|g} = 0) log f_{0,g}(x_{gj}) \Bigg| \sum_{j=1}^{m_g} \theta_{j|g} > M_g \Bigg) \\
&+ \sum_{g=1}^{G} \sum_{j=1}^{m_g} \Bigg(I(\theta_g = 0, \theta_{j|g} = 1) log f_{1,g}(x_{gj}) + I(\theta_g = 0, \theta_{j|g} = 0) log f_{0,g}(x_{gj}) \Bigg| \sum_{j=1}^{m_g} \theta_{j|g} \leq M_g \Bigg).
\end{aligned}$$

The expected value of the complete log-likelihood $l(\mathbf{x}, \mathbf{\Theta})$ with respect to latent variables given the current parameter π_1^* and $\mathbf{x}$ is

$$
\begin{aligned}
&Q(\pi_1, \pi_1^*) = E(l(\mathbf{x}, \mathbf{\Theta}|\pi_1^*, \mathbf{x})) \\
&= log(\pi_1) \sum_{g=1}^{G} P(\theta_g = 1|\pi_1^*, \mathbf{x}_g) + log(1-\pi_1) \sum_{g=1}^{G} P(\theta_g = 0|\pi_1^*, \mathbf{x}_g) \\
&+ \sum_{g=1}^{G} \sum_{j=1}^{m_g} \Big(P(\theta_{j|g} = 1|\pi_1^*, \mathbf{x}_g) log(\epsilon_g) + P(\theta_{j|g} = 0|\pi_1^*, \mathbf{x}_g) log(1-\epsilon_g) \Big) \\
&+ \sum_{g=1}^{G} \sum_{j=1}^{m_g} \Bigg(P(\theta_g = 1, \theta_{j|g} = 1|\pi_1^*, \mathbf{x}_g) log f_{1,g}(x_{gj}) \\
&+ P(\theta_g = 1, \theta_{j|g} = 0|\pi_1^*, \mathbf{x}_g) log f_{0,g}(x_{gj}) \Big| \sum_{j=1}^{m_g} \theta_{j|g} > M_g \Bigg) \\
&+ \sum_{g=1}^{G} \sum_{j=1}^{m_g} \Bigg(P(\theta_g = 0, \theta_{j|g} = 1|\pi_1^*, \mathbf{x}_g) log f_{1,g}(x_{gj}) \\
&+ P(\theta_g = 0, \theta_{j|g} = 0|\pi_1^*, \mathbf{x}_g) log f_{0,g}(x_{gj}) \Big| \sum_{j=1}^{m_g} \theta_{j|g} \leq M_g \Bigg),
\end{aligned}
$$

where

$$
\begin{aligned}
P(\theta_g = 1|\mathbf{x}_g, \pi_1^*) &= \frac{\pi_1^* P(\mathbf{x}_g|\theta_g = 1)}{\pi_1^* P(\mathbf{x}_g|\theta_g = 1) + (1-\pi_1^*) P(\mathbf{x}_g|\theta_g = 0)} \\
&= \frac{\pi_1^* \dfrac{P(\sum_{k=1}^{m_g} \theta_{k|g} > M_g | \mathbf{x}_g)}{P(\sum_{j=1}^{m_g} \theta_{j|g} > M_g)}}{\pi_1^* \dfrac{P(\sum_{k=1}^{m_g} \theta_{k|g} > M_g | \mathbf{x}_g)}{P(\sum_{j=1}^{m_g} \theta_{j|g} > M_g)} + (1-\pi_1^*) \dfrac{P(\sum_{k=1}^{m_g} \theta_{k|g} \leq M_g | \mathbf{x}_g)}{P(\sum_{j=1}^{m_g} \theta_{j|g} \leq M_g)}}, \\
P(\theta_g = 0|\mathbf{x}_g, \pi_1^*) &= \frac{(1-\pi_1^*) P(\mathbf{x}_g|\theta_g = 0)}{\pi_1^* P(\mathbf{x}_g|\theta_g = 1) + (1-\pi_1^*) P(\mathbf{x}_g|\theta_g = 0)} \\
&= \frac{(1-\pi_1^*) \dfrac{P(\sum_{k=1}^{m_g} \theta_{k|g} \leq M_g | \mathbf{x}_g)}{P(\sum_{j=1}^{m_g} \theta_{j|g} \leq M_g)}}{\pi_1^* \dfrac{P(\sum_{k=1}^{m_g} \theta_{k|g} > M_g | \mathbf{x}_g)}{P(\sum_{j=1}^{m_g} \theta_{j|g} > M_g)} + (1-\pi_1^*) \dfrac{P(\sum_{k=1}^{m_g} \theta_{k|g} \leq M_g | \mathbf{x}_g)}{P(\sum_{j=1}^{m_g} \theta_{j|g} \leq M_g)}}.
\end{aligned}
$$

To estimate π_1, we need to maximize $Q(\pi_1, \pi_1^*)$ with respect to π_1. Taking derivatives with respect to π_1 and equating them to zero, we have

$$\frac{dE(l(\mathbf{x}, \mathbf{\Theta}|\pi_1^*, \mathbf{x}))}{d\pi_1} = \frac{\sum_{g=1}^{G} P(\theta_g = 1|\pi_1^*, \mathbf{x}_g)}{\pi_1} - \frac{\sum_{g=1}^{G} P(\theta_g = 0|\pi_1^*, \mathbf{x}_g)}{1 - \pi_1} \equiv 0.$$

Finally, we can get the maximizer for π_1 as

$$\pi_1^{\text{new}} = \frac{1}{G} \sum_{g=1}^{G} P(\theta_g = 1|\pi_1^*, \mathbf{x}_g).$$

A.3 Proof of Theorem 1

In order to get (3.5), we first compute $E(\sum_{g=1}^{G} \sum_{j=1}^{m_g} \theta_{gj} | \mathbf{x})$ as follows:

$$\begin{aligned}
E\Big(\sum_{g=1}^{G} \sum_{j=1}^{m_g} \theta_{gj} \Big| \mathbf{x}_g \Big) &= E\Big(\sum_{g=1}^{G} \sum_{j=1}^{m_g} I(\theta_g = 1, \theta_{j|g} = 1) \Big| \mathbf{x}_g \Big) = \sum_{g=1}^{G} \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1 | \mathbf{x}_g) \\
&= \sum_{g=1}^{G} \sum_{j=1}^{m_g} \frac{P(\mathbf{x}_g|\theta_g = 1, \theta_{j|g} = 1) P(\theta_{j|g} = 1|\theta_g = 1) P(\theta_g = 1)}{P(\mathbf{x}_g)} \\
&= \sum_{g=1}^{G} \sum_{j=1}^{m_g} \frac{P(\mathbf{x}_g|\theta_g = 1, \theta_{j|g} = 1) P(\theta_{j|g} = 1|\theta_g = 1) P(\theta_g = 1)}{P(\mathbf{x}_g|\theta_g = 0) P(\theta_g = 0) + P(\mathbf{x}_g|\theta_g = 1) P(\theta_g = 1)} \\
&= \pi_1 \sum_{g=1}^{G} \sum_{j=1}^{m_g} \frac{P(\mathbf{x}_g|\theta_g = 1, \theta_{j|g} = 1) P(\theta_{j|g} = 1|\theta_g = 1)}{(1 - \pi_1) P(\mathbf{x}_g|\theta_g = 0) + \pi_1 P(\mathbf{x}_g|\theta_g = 1)}. \qquad \text{(A.3)}
\end{aligned}$$

The numerator of (A.3) is

$$
\begin{aligned}
&P(\mathbf{x}_g|\,\theta_{j|g}=1,\,\theta_g=1)\,P(\theta_{j|g}=1|\theta_g=1)\\
&=\sum_{\theta_{j|g}=1,\,\boldsymbol{\theta}_{k|g}\in\Omega_{g1}}\left(\prod_{k=1}^{m_g}\frac{\epsilon_g^{\theta_{k|g}}(1-\epsilon_g)^{1-\theta_{k|g}}}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}f_{\theta_{k|g}}(x_{gk})\right)\\
&=\frac{\epsilon_g f_{1,g}(x_{gj})}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}\sum_{\boldsymbol{\theta}_{k|g}\in\Omega^*_{g1}}\left(\prod_{k,k\neq j}^{m_g}\epsilon_g^{\theta_{k|g}}(1-\epsilon_g)^{1-\theta_{k|g}}\times\{(1-\theta_{k|g})f_{0,g}(x_{gk})+\theta_{k|g}f_{1,g}(x_{gk})\}\right)\\
&=\frac{\epsilon_g f_{1,g}(x_{gj})}{f(x_{gj})P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}\sum_{\boldsymbol{\theta}_{k|g}\in\Omega^*_{g1}}\left(\prod_{k,k\neq j}^{m_g}\widetilde{fdr}_{gk}^{\,1-\theta_{k|g}}(1-\widetilde{fdr}_{gk})^{\theta_{k|g}}\right)\prod_{k=1}^{m_g}f_g(x_{gk})\\
&=\frac{(1-\widetilde{fdr}_{gj})}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}\sum_{\boldsymbol{\theta}_{k|g}\in\Omega^*_{g1}}\left(\prod_{k,k\neq j}^{m_g}\widetilde{fdr}_{gk}^{\,1-\theta_{k|g}}(1-\widetilde{fdr}_{gk})^{\theta_{k|g}}\right)\prod_{k=1}^{m_g}f_g(x_{gk})\\
&\approx\frac{(1-\widetilde{fdr}_{gj})}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}\,P\Big(\sum_{k,k\neq j}^{m_g}\theta_{k|g}>M_g-1\Big|\,\mathbf{x}_g\Big)\prod_{k=1}^{m_g}f_g(x_{gk}) \qquad \text{(A.4)}\\
&\approx\frac{(1-\widetilde{fdr}_{gj})}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}\,P\Big(\sum_{k=1}^{m_g}\theta_{k|g}>M_g\Big|\,\mathbf{x}_g\Big)\prod_{k=1}^{m_g}f_g(x_{gk}). \qquad \text{(A.5)}
\end{aligned}
$$

The approximations in (A.4) is due to (3.8) and (A.5) is due to the large number of hypotheses in the group (m_g).

Based on (A.1) and (A.2), the denominator of (A.3) is

$$
\begin{aligned}
&P(\mathbf{x}_g)=(1-\pi_1)P(\mathbf{x}_g|\theta_g=0)+\pi_1P(\mathbf{x}_g|\theta_g=1)\\
&\approx\frac{(1-\pi_1)\prod_{k=1}^{m_g}f_g(x_{gk})}{P(\sum_{j=1}^{m_g}\theta_{j|g}\leq M_g)}P(\sum_{k=1}^{m_g}\theta_{k|g}\leq M_g|\,\mathbf{x}_g)+\frac{\pi_1\prod_{k=1}^{m_g}f_g(x_{gk})}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}\,P(\sum_{k=1}^{m_g}\theta_{k|g}>M_g|\,\mathbf{x}_g).
\end{aligned}
\qquad \text{(A.6)}
$$

Therefore, by (A.5) and (A.6), (A.3) is rewritten as follows:

$$
\begin{aligned}
E\Big(\sum_{g=1}^{G}\sum_{j=1}^{m_g}\theta_{gj}\Big|\,\mathbf{x}_g\Big) &= \pi_1\sum_{g=1}^{G}\sum_{j=1}^{m_g}\frac{P(\mathbf{x}_g|\theta_g=1,\theta_{j|g}=1)P(\theta_{j|g}=1|\theta_g=1)}{P(\mathbf{x}_g)} \\
&= \pi_1\sum_{g=1}^{G}\sum_{j=1}^{m_g}\frac{\dfrac{(1-\widetilde{fdr}_{gj})}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}\,P(\sum_{k=1}^{m_g}\theta_{k|g}>M_g|\,\mathbf{x}_g)\prod_{k=1}^{m_g}f_g(x_{gk})}{\dfrac{(1-\pi_1)\prod_{k=1}^{m_g}f_g(x_{gk})}{P(\sum_{j=1}^{m_g}\theta_{j|g}\le M_g)}P(\sum_{k=1}^{m_g}\theta_{k|g}\le M_g|\,\mathbf{x}_g)+\dfrac{\pi_1\prod_{k=1}^{m_g}f_g(x_{gk})}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}\,P(\sum_{k=1}^{m_g}\theta_{k|g}>M_g|\,\mathbf{x}_g)} \\
&= \sum_{g=1}^{G}\frac{\dfrac{\pi_1P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}}{\dfrac{(1-\pi_1)P(\sum_{j=1}^{m_g}\theta_{j|g}\le M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g}\le M_g)}+\dfrac{\pi_1P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}}\sum_{j=1}^{m_g}(1-\widetilde{fdr}_{gj}) \\
&= \sum_{g=1}^{G}w_g\sum_{j=1}^{m_g}(1-\widetilde{fdr}_{gj}),
\end{aligned}
$$

where $\widetilde{fdr}_{gj}$ is defined in (3.7) and

$$
w_g=\frac{\dfrac{\pi_1P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}}{\dfrac{(1-\pi_1)P(\sum_{j=1}^{m_g}\theta_{j|g}\le M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g}\le M_g)}+\dfrac{\pi_1P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g|\,\mathbf{x}_g)}{P(\sum_{j=1}^{m_g}\theta_{j|g}>M_g)}}.
$$

Thus, (3.5) can be expressed as

$$
\begin{aligned}
E\Big(\frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}(1-\theta_{gj})\delta_{gj}(\mathbf{x})}{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\mathbf{x})\vee 1}\Big|\,\mathbf{x}\Big) &= 1-\ E\Big(\frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\theta_{gj}\delta_{gj}(\mathbf{x})}{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\mathbf{x})\vee 1}\Big|\,\mathbf{x}\Big) \\
&= 1-\frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}E(\theta_{gj}|\mathbf{x})\delta_{gj}(\mathbf{x})}{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\mathbf{x})\vee 1} \\
&= 1-\frac{\sum_{g=1}^{G}\delta_g(\mathbf{x}_g)w_g\sum_{j=1}^{m_g}(1-\widetilde{fdr}_{gj})\delta_{j|g}(\mathbf{x}_g)}{\sum_{g=1}^{G}\delta_g(\mathbf{x}_g)\{\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g)\}\vee 1} \\
&= 1-\frac{\sum_{g=1}^{G}w_g\delta_g(\mathbf{x}_g)\{I(\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g))-PFDR_{w|g}(\mathbf{x}_g)\}\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g)}{\sum_{g=1}^{G}\delta_g(\mathbf{x}_g)\{\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g)\}\vee 1} \\
&= \frac{\sum_{g=1}^{G}\delta_g(\mathbf{x}_g)[1-w_g\{I(\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g))-PFDR_{w|g}(\mathbf{x}_g)\}]\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g)}{\sum_{g=1}^{G}\delta_g(\mathbf{x}_g)\{\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g)\}\vee 1},
\end{aligned}
$$

where

$$
PFDR_{w|g}(\mathbf{x}_g)=\frac{\sum_{j=1}^{m_g}\widetilde{fdr}_{gj}\delta_{j|g}(\mathbf{x}_g)}{\sum_{j=1}^{m_g}\delta_{j|g}(\mathbf{x}_g)\vee 1}.
$$

Bibliography

Pan-ukb team, 2020. URL `https://pan.ukbb.broadinstitute.org`.

Turki A Althunian, Anthonius de Boer, Rolf HH Groenwold, and Olaf H Klungel. Defining the noninferiority margin and analysing noninferiority: an overview. *British Journal of Clinical Pharmacology*, 83(8):1636–1642, 2017.

R Foygel Barber and Aaditya Ramdas. The p-filter: Multi-layer fdr control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Afiat Berbudi, Nofri Rahmadika, Adi Imam Tjahjadi, and Rovina Ruslami. Type 2 diabetes and its impact on the immune system. *Current diabetes reviews*, 16(5): 442, 2020.

James O Berger and Mohan Delampady. Testing precise hypotheses. *Statistical Science*, pages 317–335, 1987.

Colin R Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

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

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

Shun-Yi Chen and Hubert J Chen. Range test for the equivalency of means under unequal variances. *Technometrics*, 41(3):250–260, 1999.

Sungwoo Choi and Junyong Park. Plug-in tests for nonequivalence of means of independent normal populations. *Biometrical Journal*, 56(6):1016–1034, 2014.

Miranda E Cox, Joel K Campbell, and Carl D Langefeld. An exploration of sex-specific linkage disequilibrium on chromosome x in caucasians from the coga study. 6(1):1–4, 2005.

Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801, 2018.

Paola De Candia, Francesco Prattichizzo, Silvia Garavelli, Veronica De Rosa, Mario Galgani, Francesca Di Rella, Maria Immacolata Spagnuolo, Alessandra Colamatteo, Clorinda Fusco, Teresa Micillo, et al. Type 2 diabetes: how much of an autoimmune disease? *Frontiers in Endocrinology*, 10:451, 2019.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228, 2015.

RA Fisher. Statistical methods for research workers (london: Oliver and boyd). *Legends to Figures*, 1932.

Steven Gazal, Hilary K Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M Neale, Alexander Gusev, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10): 1421, 2017.

Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

Hector Giral, Ulf Landmesser, and Adelheid Kratzer. Into the wild: Gwas exploration of non-coding rnas. *Frontiers in cardiovascular medicine*, 5:181, 2018.

Ruth Heller, Nilanjan Chatterjee, Abba Krieger, and Jianxin Shi. Post-selection inference following aggregate level hypothesis testing in large-scale genomic data. *Journal of the American Statistical Association*, 113(524):1770–1783, 2018.

Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Nan M Laird and Christoph Lange. *The fundamentals of modern statistical genetics*. Springer Science & Business Media, 2010.

EL Lehmann and Joseph P Romano. Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138–1154, 2005.

Yanping Liu, Sanat K Sarkar, and Zhigen Zhao. A new approach to multiple testing of grouped hypotheses. *Journal of Statistical Planning and Inference*, 179:1–14, 2016.

Piero Marchetti, Marco Bugliani, Vincenzo De Tata, Mara Suleiman, and Lorella Marselli. Pancreatic beta cell identity in humans and the role of type 2 diabetes. *Frontiers in cell and developmental biology*, 5:55, 2017a.

Piero Marchetti, Marco Bugliani, Vincenzo De Tata, Mara Suleiman, and Lorella Marselli. Pancreatic beta cell identity in humans and the role of type 2 diabetes. *Frontiers in cell and developmental biology*, 5:55, 2017b.

Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Athanasios Teodosiadis, et al. Index and biological spectrum of human dnase i hypersensitive sites. *Nature*, 584(7820):244–251, 2020.

Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, Rajinder Kaul, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

Debmalya Sanyal. Diabetes is predominantly an intestinal disease. *Indian journal of endocrinology and metabolism*, 17(Suppl1):S64, 2013.

Sanat K Sarkar, Aiying Chen, Li He, and Wenge Guo. Group sequential bh and its adaptive versions controlling the fdr. *Journal of Statistical Planning and Inference*, 199:219–235, 2019.

Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

Elizabeth K Speliotes, Cristen J Willer, Sonja I Berndt, Keri L Monda, Gudmar Thorleifsson, Anne U Jackson, Hana Lango Allen, Cecilia M Lindgren, Reedik Mägi, Joshua C Randall, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, 42(11):937–948, 2010.

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

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

John D Storey, James Y Dai, and Jeffrey T Leek. The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, 8(2):414–432, 2007.

Wenguang Sun, Brian J Reich, T Tony Cai, Michele Guindani, and Armin Schwartzman. False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):59–83, 2015.

Roy Taylor. Insulin resistance and type 2 diabetes. *Diabetes*, 61(4):778–779, 2012.

John Wilder Tukey. The problem of multiple comparisons. *Multiple comparisons*, 1953.

Jon Wakefield. Reporting and interpretation in genome-wide association studies. *International Journal of Epidemiology*, 37(3):641–653, 2008.

Brian L Wiens. Choosing an equivalence limit for noninferiority or equivalence studies. *Controlled clinical trials*, 23(1):2–14, 2002.