Tests for detection of rare variants and gene-environment interaction in cohort and twin family studies

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Brandon J Coombes

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advised by Saonli Basu

August, 2016

Acknowledgments

I am very much indebted to my advisor, Saonli Basu, for guiding me from start to finish through my dissertation. Special thanks to my committee, Jim Hodges, Eric Lock, and Matt McGue, for their helpful suggestions and feedback. I would also like to thank Saonli Basu, Matt McGue, Mike Miller, and Jaron Arbet for their invaluable feedback during lab meetings. I would like to acknowledge the UMN graduate school for funding me through the Doctoral Dissertation Fellowship. I would also like to thank Dr. Wei Pan and the National Institute of General Medical Sciences for training me as a Statistical Genetics Training Grant Fellow. Finally, I would like to thank my friends and family for enduring with me through this part of my life. Most importantly, I would be lost without my wife, Dr. Courtney Coombes. It was incredible to be able to both go through our respective PhD programs at the same time. She was a great encourager and proof-reader.

Dedication

To Courtney for putting up with me.

Abstract

Complex diseases are caused by a combination of environmental and genetic factors. While we have estimated that genetic factors explain a large proportion of variance in many of these diseases, current strategies using only the genotyped common variants (CVs) have failed to explain all of this heritability. There are many hypotheses for this so-called "missing heritability." We study two such hypotheses by extending a sequential algorithm that was initially proposed to test for genetic main effects of a candidate gene. We first extend the model selection test using the sequential algorithm to a model-averaging test. We use these tests to study how rare variants (RVs) rather than CVs may explain a larger proportion of the disease risk and apply our methods to a candidate gene study of obesity that has sequenced CVs and RVs.

It is also thought that the effect of the variants is moderated by environmental factors. Thus, gene-environment interactions may explain why we are not able to identify genes that cause disease. To improve power to detect gene-environment interactions for variants within a candidate gene in studies with unrelated or related subjects, we extend the sequential algorithm for the model selection test for genetic main effects to instead test for these interactions. For studies with unrelated subjects, we extend the sequential algorithm to create summary measures for either the genetic main effect or the interaction and show that these tests are often valid under realistic scenarios. We use a combination of the main effect and interaction summary measures to powerfully test for gene-environment interaction in a variety of situations. We apply our method to test whether candidate genes interact with family climate to influence alcohol consumption among a parent population.

Lastly, we extend the tests of gene-environment interaction for unrelated subjects to families. We model the family data using a linear mixed model (LMM) framework to

account for shared genetic and environmental effects within a family. In order to reduce the number of parameters we need to estimate, we propose using a ridge penalty on the genetic main effect re-expressed as a random effect within the LMM. We also develop a test which is weighted version of a previous test using the sum of powers of the score vector for interaction where weights are chosen with our sequential algorithm. We show that this test can be more powerful than the previous test when there are a mix of positive and negative interaction effects. We apply our test to a twin study to identify significant interactions between the CVs of candidate genes and a set of environmental factors that influence alcohol consumption.

Contents

Acknowledgments i

Dedication ii

Abstract iii

List of Tables viii

List of Figures ix

1 Introduction 1

1.1 Overview 1

1.2 Rare variants 2

1.3 GxE interaction 2

1.4 Sequential algorithm to test for effect 3

1.5 Dissertation objectives 3

2 Weighted score tests implementing model-averaging schemes in detection of rare variants in case-control studies 6

2.1 Introduction 6

2.2 Materials and methods 9

2.2.1 Existing approaches 10

2.2.2 Model branching through thresholding 13

2.2.3 Selection of models using a weighted likelihood function 14

2.2.4 A weighted score test 15

2.2.5 Paring the branches . . . 16
2.3 Results . . . 16
2.3.1 Simulation study . . . 16
2.3.2 Simulation 1: Null distribution of the test statistic . . . 17
2.3.3 Comparing power among different approaches . . . 18
2.3.4 Sanofi Data . . . 23
2.4 Discussion . . . 25

3 A combination test for detection of gene-environment interaction in cohort studies 33
3.1 Introduction . . . 33
3.2 Methods . . . 35
3.2.1 MinP test . . . 36
3.2.2 Score tests of interaction . . . 37
3.2.3 G-E interaction using summary measures . . . 38
3.3 Results . . . 43
3.4 Minnesota Center for Twin and Family Research . . . 51
3.5 Discussion . . . 53

4 Score tests for gene-environment interaction in family studies using linear mixed models 56
4.1 Introduction . . . 56
4.2 Methods . . . 59
4.2.1 Joint modeling of GxE interaction . . . 59
4.2.2 Univariate modeling of GxE interaction . . . 65
4.3 Results . . . 65
4.3.1 Simulation 1 . . . 67
4.3.2 Simulation 2 . . . 69
4.3.3 Simulation 3 . . . 69
4.4 Minnesota Center for Twin and Family Research . . . 70
4.5 Discussion . . . 72

5 Conclusion and Discussion **78**
5.1 Methodological advances 78
5.2 Future directions for research 80

References **82**

Appendix A. Supplementary materials for Chapter 3 **91**
A.1 Proof of asymptotic distribution of iSeq-aSum-I 91
A.2 Correlation between burden summary measure and environment 93
A.3 Performance of minP test for interaction testing 94

Appendix B. Supplementary materials for Chapter 4 **96**
B.1 Ridge regression as a random effect 96
B.2 AE model estimated covariance when the true model is ACE 97
B.3 Computation time 97

List of Tables

2.1 Effect of order dependency on model selection 19
2.2 Power Comparison in presence of no LD 20
2.3 Power Comparison in presence of LD 22
2.4 Power comparison with a mix of CVs and RVs 29
2.5 Top p-values of Sanofi data 32
3.1 Genes that interact with family climate in MCTFR parents 53
4.1 ACE versus AE model comparison 68
4.2 Comparison of ridge penalty versus fixed effect modeling 74
4.3 Large differences in SPU for MCTFR data 77
B.1 Mean computation time 98

List of Figures

2.1 A demonstration of benefit of model-averaging compared to model selection 12
2.2 Null distributions of model selection and model-averaging 28
2.3 Sanofi analysis of FAAH gene 30
2.4 Sanofi analysis of MGLL gene 31
3.1 LD structure of ADH1B and ALDH1A1 45
3.2 Type I error of the G-E interaction methods as the amount of G-E dependence is increased 47
3.3 Comparison of power for G-E interaction methods for independent subjects 49
3.4 Effect of genetic main effect on G-E interaction methods 50
4.1 Comparison of SPU and Seq-SPU 75
4.2 MCTFR twin analysis stratified by sex 76

Chapter 1

Introduction

1.1 Overview

Human diseases can result from a combination of environmental and genetic factors. Diseases caused by genetic factors are said to be heritable because they are transmissible from parent to progeny. We can estimate the heritability of a disease by examining what proportion of observed differences in disease is due to genetic differences (Wray and Visscher, 2008). Heritability provides us with an estimate of how strongly a trait or disease can be passed from generation to generation. Some heritable disorders can easily be explained by simple genetics where the presence of a certain gene type means almost certain presence of the disorder. Genes of these Mendelian inherited diseases have been successfully identified by genetic studies. However, some diseases known to be heritable, such as schizophrenia, Alzheimer's, substance use disorders (SUDs), and obesity, and cannot be explained purely by genetics. In these complex diseases, simply having a certain genotype does not always cause the disorder, yet it does seem to be passed down from parent to offspring.

A common study design to investigate which genes contribute to a complex disease's heritability is called a genome-wide association study (GWAS) (Manolio, 2010). GWASs collect genotype data on a huge number, usually in the millions, of single-nucleotide polymorphisms (SNPs). A SNP is a common variant in a specific location in the DNA. Large groups of people share the same SNPs because DNA has been passed down from parent to offspring for many generations. Using a phenotype for a given

disease and the genotype data, GWAS aim to identify SNPs that are associated with the disease. While many new variants associated with disease have been identified by GWAS, for many diseases, these variants still explain very little of the genetic risk or heritability. There are several hypotheses to explain this "missing heritability" (Manolio et al. (2009), Zuk et al. (2012), Lee et al. (2011), Slatkin (2009)). In this dissertation, we are interested in investigating two such hypotheses: rare variants and gene-environment (GxE) interaction.

1.2 Rare variants

One hypothesis to explain the missing heritability of a disease is that rare variants (RVs) in the DNA contribute to disease. Rare variants are mutations in the DNA that occur in a few generations or maybe even a single individual. Because of their rarity, these low frequency variants could have substantial effect sizes without demonstrating clear Mendelian inheritance and could contribute substantially to missing heritability (Schork et al., 2009) However, due to the low minor allele frequency of an RV, it is difficult to detect individual effects, and it is impossible to model if a RV is unique to one individual. Thus, considering a group of RVs is necessary for any RV analysis. In Chapter 2, we aim to explore these variations, common (CVs) and rare (RVs), to explain disease and develop strategies for analyzing them.

1.3 GxE interaction

Another hypothesis to explain the missing heritability of a disease is that genetic factors associated with disease are missed because their effects are being moderated by environmental factors (Kaprio, 2012). Genes and environments can interact in many ways (Ottman, 1990): a genotype may increase expression of an environmental risk factor, a genotype can exacerbate the environmental effect on disease or vice versa, a genotype and environmental risk factor could both be required for disease, or gene and environment each influence disease risk by themselves. The identification of main effect risk factors, either genetic or environmental, represents only one of the components that contribute to complex diseases, so studying GxE interactions can increase

the chances of detecting genes with only a small marginal effect in disease development. Studying these interactions would also reveal risk factors specific to certain etiological backgrounds that might otherwise be undetectable because their effects are evident only in subgroups of subjects and not for subjects considered as a whole. Consequently, identifying GxE interactions could lead to better diagnostics of disease and eventually personalized medicine or therapies. In Chapters 3 and 4, we develop methods to study GxE interaction in a study of independent subjects or families, respectively.

1.4 Sequential algorithm to test for effect

The high-dimensional aspect of genetic data makes RVs and GxE interactions challenging to model. In a given study, there may be millions of genetic variants to consider. Additionally, while narrowing the analysis to a candidate gene reduces dimensionality and allows for easier interpretation, there may still be thousands of variants in the analysis. In order to give a solution for either of these hypotheses of missing heritability, we aim to reduce dimensionality. While there are many strategies to reduce dimension, we use the sequential algorithm proposed by Basu and Pan (2011). This algorithm aims to powerfully combine effects through an optimal model selected by searching over different models using a stepwise procedure. In Chapter 2, we propose a model-averaging extension to this algorithm to better search the model space when analyzing the genetic effects of SNPs, RVs, or a combination. In Chapters 3 and 4, we extend this algorithm in different ways to test for GxE interaction in studies with independent subjects or families, respectively.

1.5 Dissertation objectives

The remainder of the dissertation is laid out as follows. In Chapter 2, we develop a model-averaging extension of the sequential algorithm of Basu and Pan (2011) to test for genetic effects for a set of variants from a candidate gene. The model selection method of Basu and Pan (2011) searches how to best sum the genetic main effect. It does this by initially summing the minor alleles of all of the variants. It then sequentially decides, using the likelihood, whether the minor alleles for the current variant should

be left out of the sum or instead subtracted from the sum. However, if two likelihoods are close, it may be better to explore both options for how to sum variants. Moreover, selection of the optimal model using this algorithm can depend on the order of the genetic variants. We instead propose exploring both options when the likelihoods are close at a given step. This creates a branching-type process that can search more models than its predecessor. After this process is complete, we average the models together using a weighted score test. While the advantages of model-averaging have been well documented in the prediction literature (Viallefont et al., 2001), we study the advantage of model-averaging over model selection when our purpose is inference. Using the Sanofi case-control dataset (Bansal et al., 2011), we apply model selection and model-averaging to test whether the SNPs and RVs for two candidate genes are associated with obesity.

In Chapter 3, we test for interactions between a candidate gene and an environmental factor while improving power by pooling multiple variants within a gene. We extend recently developed testing approaches based on score statistics (Pan et al., 2014) to the gene-environment interaction-testing problem. We also propose tests for interactions using gene-based summary statistics, including one produced by the sequential algorithm of Basu and Pan (2011). While it has recently been shown that these summary measures can be biased and may lead to inflated type I error (Lin et al., 2015), we show that under several realistic scenarios, we can still provide valid tests of interaction. These tests use significantly fewer degrees-of-freedom and thus can have much higher power to detect interaction. Additionally, we demonstrate that a combination test using summary measures provides a powerful alternative for testing for gene-environment interaction. We demonstrate the usefulness of these approaches using simulation studies and illustrate their performance by studying the interaction between the SNPs in several candidate genes and family climate environment on alcohol consumption using the Minnesota Center for Twin and Family Research dataset (McGue et al., 2014).

In Chapter 4, we use a linear mixed model to test for interaction between a set of correlated environments and the variants in a candidate gene for family studies. The environments can either be modeled in independent models or jointly in one model. In our simulations, we find that the joint model is typically best even if only one environment has interaction with the gene. For either strategy, we also propose treating the genetic

main effect in this model as a random effect in order to reduce the number of parameters to capture the main effect. Using the score vector of the gene-environment interactions and its covariance, we are able to extend many current tests of gene-environment interaction in candidate genes to family data. Additionally, we propose a generalization of the test of interaction from Chapter 3 that adaptively sums the interactions using the sequential algorithm of (Basu and Pan, 2011). This generalized set of tests, referred to as the Seq-SPU family of tests, can be expressed as a weighted version of the sum of power score tests (SPU) (Pan et al., 2014; Kim et al., 2014). We find that the adaptive version of our test, Seq-aSPU, can outperform aSPU in cases where the interaction effects are a mix of positive and negative and few null interactions. We applied these methods to the Minnesota Center for Twin and Family Research dataset and found one significant gene that interacts with four psychosocial environmental factors affecting the alcohol consumption among the twins. This significant gene was only identified by our test, Seq-aSPU.

Finally, we conclude the dissertation with a discussion in Chapter 5.

Chapter 2

Weighted score tests implementing model-averaging schemes in detection of rare variants in case-control studies

2.1 Introduction

Genome-wide association studies (GWASs) have successfully identified many common genetic variants that are associated with a given outcome, but little risk can be explained by these identified single nucleotide polymorphisms (SNPs). There are several hypotheses for genetic factors contributing to disease risk (Manolio et al., 2009; Zuk et al., 2012; Lee et al., 2011; Slatkin, 2009). One such hypothesis is that rare variants (RVs) measured in sequencing studies with large effect sizes contribute to the disease risk. However, the low minor allele frequency (MAF) of a RV makes it difficult to detect individual effects. Thus, rare-variant models are used to detect the combined effect of a set of RVs, such as the RVs within a candidate gene.

The existing approaches for rare variant detection can be broadly classified into three separate categories: (1) Collapsing methods based on pooling multiple RVs such as the Sum test (Pan, 2009), Cohort Allelic Sums Test (CAST) (Morgenthaler and Thilly,

2007), Combined Multivariate and Collapsing (CMC) (Li and Leal, 2009), Weighted Sum (W-Sum) test of Madsen and Browning (2009), Kernel Based Adaptive Cluster (KBAC) (Liu and Leal, 2010), Replication Based Test (RBT) (Ionita-Laza et al., 2011), ARIEL test (Asimit et al., 2012), and the EREC test (Lin and Tang, 2012); (2) methods based on model selection such as Seq-aSum and Seq-aSum-VS approaches (Basu et al., 2011; Basu and Pan, 2011), Variable Threshold Test (VT) (Price et al., 2010), RARECOVER method (Bhatia et al., 2010), Selective grouping method (Zhang et al., 2011), and Step-Up approach (Hoffmann et al., 2010); and (3) methods based on treating RV effects as random effects such as the SSU approach (Pan, 2009), C-alpha test (Neale et al., 2011), and SKAT approach (Wu et al., 2011). Basu and Pan (2011) studied the performance of several of these multi-marker tests under a variety of disease models. The Sum test (Pan, 2009) was most powerful when there were no causal variants with effects in opposite directions and when there were few or no non-causal RVs; otherwise, it suffered from substantial loss of power. In the presence of opposite association directions and non-causal RVs, the SSU and SKAT tests performed better than the other tests. The model-selection approaches performed in the middle of random effect and collapsing methods. According to Basu and Pan (2011), the model selection methods, especially Seq-aSum-VS approach, performed very well when there were both protective and deleterious causal RVs and very few non-causal RVs, but the performance of the Seq-aSum-VS approach was not very impressive in the presence of a moderate or large number of non-causal RVs. These and other findings (Basu and Pan, 2011) have led to combining the strengths of collapsing and random effect methods such as SKAT-O (Lee et al., 2012), Fisher method (Derkach et al., 2013) and MiST (Sun et al., 2013) as discussed in a recent review (Lee et al., 2014). Also, it was recently suggested that using SKAT in the presence of RVs and common variants (CVs) may be less optimal because RVs are weighted to have much more importance than CVs (Ionita-Laza et al., 2013). To overcome this, an upweighting of the CVs was implemented in SKAT-C.

While many improvements have been made in the random effects and collapsing methods, this paper takes a closer look at the methods based on model selection, especially the Seq-aSum and Seq-aSum-VS approaches. The Seq-aSum-VS approach classifies RVs based on the direction of association ('+1' for positive association, '-1' for negative association and '0' for no association) and implements a sequential variable

selection scheme to select the best model for association between the SNP-set and the disease. The only difference between the Seq-aSum approach and the Seq-aSum-VS approach is that the variable selection ('0' allocation for a variant) is not implemented in the former. The Seq-aSum-VS approach starts with putting all the RVs in the '+1' group and proceeds by moving each RV sequentially to the other two groups and finally chooses the allocation ('+1','-1', or '0') with highest likelihood to the RV. The process of choosing the best model in Basu and Pan (2011)'s method can be compared to a stepwise regression, where one may not always find the best model due to this selection scheme. This is especially true if a particular allocation results in a slightly higher likelihood than the other two allocations. In this case, choosing the allocation with highest likelihood for a SNP might not be optimal, rather it might be more efficient to allow multiple allocations for a RV and construct a test that takes into account multiple plausible models for the disease-RV association. Moreover, the performance of the sequential search often depends on the ordering of the variants in this search mechanism. A model-averaging approach could potentially reduce the dependency on the ordering of the variants in this sequential search.

Another issue to note here is that model selection approaches use dimension-reduction strategies to substantially reduce the number of parameters one would require to fit these large number of RVs. Hence, any model we can construct will never be the true model that generated the data we observe. In other words, the set of models is clearly misspecified, and model selection is best seen as a way of approximating, rather than identifying, full reality (Burnham and Anderson (2002), pp. 20-23). A model-averaging approach, on the other hand, could have an advantage over this model selection scheme. By averaging over a number of models, a model-averaging approach reduces the uncertainties associated with selection of models. However, averaging over a large number of models, especially the uninformative ones, could cause loss in power. In addition, the approach could be too computationally intensive to be useful.

Model-averaging has been well studied in the prediction literature. One popular method is Bayesian model-averaging (BMA) (Raftery et al., 1997; Hoeting et al., 1999; Viallefont et al., 2001) which provides guidelines for selecting a subset of plausible models, calculates the posterior probability that a predictor has an effect given disease status, and improved predictive performance of this model-averaging approach over a

model selection approach. However, there is not much literature on the performance of model-averaging in testing the effect of a predictor. Moreover, due to the low occurrence of RVs within the case-control set, RVs would have to be considered in aggregate, which complicates implementation of a model-averaging scheme. Here we aim to compare the performance of a model-averaging approach in detection of RVs versus a model selection approach, where the models that contribute to the model-averaging approach are selected in a data-adaptive way.

This work proposes a data-adaptive model-averaging technique that addresses the limitation in the Seq-aSum-VS approach. Specifically, we allow selection of a set of potential models through our model selection scheme and use a weighted score test to detect association instead of choosing the best model. The rest of the paper is organized as follows. In the Methods section, we describe the several existing approaches and propose several alternative model-averaging schemes. In the Results section, we compare the proposed schemes with model selection approaches through extensive simulation studies and a real data example. We conclude with a short summary and discussion outlining a few future research topics.

2.2 Materials and methods

The purpose of this study is to develop methods to improve the power for detecting an association between a trait and a group of RVs, for example, RVs in a sliding window or in a functional unit such as a gene. Although we have only considered binary traits here, our method and some of the other methods can be easily extended to other types of traits.

Assume there are n unrelated individuals, where $Y_i = 0$ for n_0 controls and $Y_i = 1$ for n_1 cases and $n_0 + n_1 = n$. The k RVs are denoted by $X_{ij}, j = 1, ..., k$ for $i = 1, 2, \ldots, n$. The variables X_{ij} can take values such as 0, 1 and 2 corresponding to the number of minor alleles present. We do not take into account adjusting for covariates, such as environmental factors, though all of the methods are based on logistic regression and can accommodate covariates.

2.2.1 Existing approaches

A logistic main effect model to test for association between a binary trait Y_i and k RVs is written as

$$\text{Logit}\Pr(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j; \quad i = 1, ..., n. \tag{2.1}$$

To test for association between the trait and these k RVs, the null hypothesis of no association is formulated as

$$H_0 : \boldsymbol{\beta} = (\beta_1, ..., \beta_k)' = 0.$$

One could perform a score test for the null hypothesis H_0, but estimating the effect of an individual RV may not be feasible and the approach loses power in the presence of many null RVs. Hence, different techniques such as collapsing methods, random effect models, and model selection methods mentioned in the introduction have been proposed to handle these high-dimensional RVs.

Basu et al. (2011) and Basu and Pan (2011) have proposed the Seq-aSum and Seq-aSum-VS approaches to incorporate model selection in constructing a test for association. These approaches attempt to sort the SNPs/RVs into one of three groups (null, causal, or protective). The general model for these approaches is based on the model suggested by Hoffmann et al. (2010),

$$\text{Logit}\Pr(Y_i = 1) = \beta_0 + \beta_c \sum_{j=1}^{k} X_{ij}\gamma_j; \quad i = 1, ..., n; \quad j = 1, ..., k, \tag{2.2}$$

with $\gamma_j = w_j s_j$, where w_j is a weight assigned to RV j, $s_j = 1$ or -1 indicating whether the effect of RV j is positive or negative, and $s_j = 0$ indicating the exclusion of RV j from the model (i.e., the SNP is unlikely to be associated with the trait). There is literature on how to choose appropriate w_j for a specific problem and, if needed, it is not difficult to incorporate such weights into the methods we describe in this section. Here we assume $w_j = 1$ for all $j = 1, 2, \ldots, k$ in any subsequent analysis.

The null hypothesis for testing the association between the trait and the RVs boils down to $H_0 : \beta_c = 0$. Seq-aSum and Seq-aSum-VS use the data to adaptively determine the optimal allocation. The Seq-aSum-VS test proceeds through the following sequence of events:

1. Start with $s_j = 1$ for all j.
2. `for` j in $1 : k$
   (a) Compute the maximized likelihood (maximized over β_c) corresponding to $s_j = -1, 0, 1$.
   (b) Set s_j to the value that corresponds to the largest maximized likelihood among the three possible allocations (-1, 0, or 1).

The method selects a model from among $2k+1$ candidate models. Due to the sequential nature of the estimation, it is not guaranteed that the best model (model with the highest likelihood) will be chosen. Nevertheless, it avoids searching over 3^k possible models and thus gains power in many situations, especially for a large number of RVs.

The Seq-aSum-VS approach chooses the best model among $2k + 1$ models and thus allows for only one allocation (s_j = 0, 1, or -1) for each RV j, $j = 1, 2, \ldots, k$. For a large number of neutral RVs, choosing the best model might not be an efficient way to detect association. A neutral RV j does not necessarily give highest likelihood at $s_j = 0$. For a given dataset, it could have a non-significant increase in the likelihood at allocation $s_j = 1$ or $s_j = -1$. Choosing the allocation that provides highest likelihood for such a SNP could affect the optimal assignment of the following RVs. It might be more efficient to allow multiple allocations for a RV instead of choosing the one with the highest likelihood and to construct a test that takes into account multiple plausible models for the disease-RV association.

One such scenario where the model-averaging approach has an advantage over the model-selection approach is where only the last RV is causal. Here, a sequential model selection algorithm could fail to find the best allocation due to the null RVs diluting the effect of the lone causal RV (Basu and Pan, 2011). For demonstration, we consider a scenario where out of 4 independent RVs in a set, only the last RV is causal. Fig 2.1 shows the paths taken for model selection (top) and model-averaging (bottom) for one such realization. By construction, model selection selects only one path. Model-averaging, however, computes how likely a given path is at each node and explores all likely paths. Section 2.2.2 explains one way to construct a measure to choose "likely" paths. In Fig 2.1, the model-averaging algorithm finds a better path (the bottom path) than the one selected by model selection. By averaging the four likely paths, we reduce

the dependency on ordering and thus gain power to detect association of the RVs with disease. In the next two subsections, we propose two path-finding algorithms to identify potential models to capture association between k RVs and the binary disease.

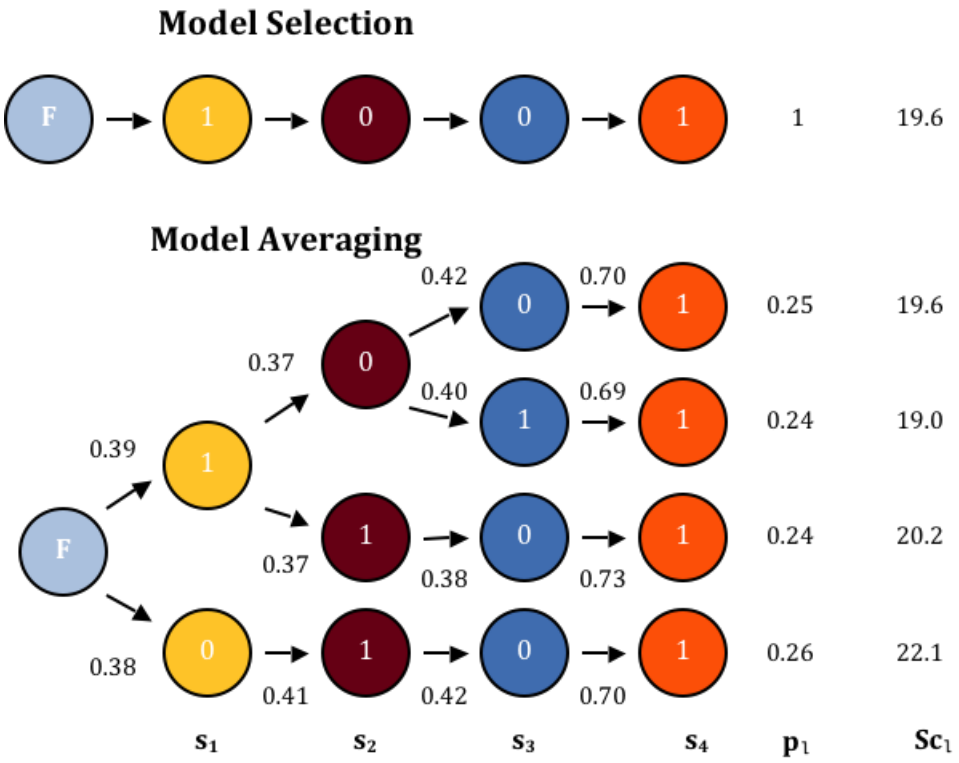

Figure 2.1: A demonstration of benefit of model-averaging (bottom) compared to model selection (top). The simulation setup is described in Section 2.3.3. F denotes the start of each algorithm where all RVs are allocated to the 'positive' group ($s_j = 1$ for all j). The numbers above the directional arrows on the model-averaging figure are ratios indicating how likely a given allocation at the current step is selected. The paths with the highest ratio and any "close" ratios are explored. p_l is the path weight given by multiplying ratios along a path and scaling them so that the sum of path weights equals 1. Sc_l is the score statistic for each path (See Section 2.2.2 for details).

2.2.2 Model branching through thresholding

This proposed approach provides a way to select multiple possible allocations ('+1','-1', or '0') for a RV. This method proceeds the same way as the Seq-aSum-VS approach. It starts with putting all the RVs in '+1' group and proceeds by moving each RV sequentially to the other two groups ('0' or '-1' group). The Seq-aSum-VS approach classifies the RV to the allocation with the highest likelihood, but the proposed approach allows multiple plausible allocations of a RV by selecting allocations that give high scores rather than just the highest score. If any two allocation scores are close, we explore both paths. The tree model proposed here proceeds to construct a tree in the following way:

1. Set $s_j = 1$ for $j = 1, ..., k$.

2. Choose a cutoff, κ, between 0 and 1 that determines if a score statistic is close to the highest score statistic.

3. Starting with s_1, calculate the three score statistics corresponding to setting $s_1 = 1, 0$ and -1. Denote them as Sc_1, Sc_0, and Sc_{-1}, respectively. The score statistic Sc_l is given by

$$Sc_l = \frac{\left[\sum_{i=1}^{n}(Y_i - \bar{Y})(g_{i,l} - \bar{g}_l)\right]^2}{\bar{Y}(1-\bar{Y})\sum_{i=1}^{n} g_{i,l}^2},$$

   where $g_{i,l} = \sum_{j=1}^{k} X_{ij} s_l$ with $s_l = 0, 1, \text{or} - 1$ and $s_{l'} = 1$ for all $l' \neq l$

4. Calculate the score ratio

$$R_l = \frac{Sc_l}{Sc_1 + Sc_0 + Sc_{-1}} \tag{2.3}$$

   for $l = 1, 0$, and -1.

5. Allow allocations satisfying $R_l \geq \max_l\{R_l\} \times \kappa$, similar to the "Occam's window" technique proposed by Madigan and Raftery (1994)

6. For each allowed allocation of s_1,
   `for` $j = 2 : k$

7. Repeat Steps 3-5 to find possible values for $s_j,\ j = 2, \ldots, k$.

8. Finally, obtain a tree, each branch of which represents a possible allocation of $(s_1, ..., s_k)$.

9. For convenience, the weight of each branch p_l is calculated by taking the product of score ratios of successive branching steps.

Let there be finally m total branches, with overall weights $p_1, p_2, ..., p_m$ respectively. We propose a weighted score test using the above data adaptive model-averaging approach to test for the null hypothesis H_0 in Section 2.2.4. We refer to this approach as Branching under Ratios (BUR) approach.

2.2.3 Selection of models using a weighted likelihood function

We also propose a model-averaging approach through a weighted likelihood function. To allow multiple plausible allocations for a RV, we assume

$$s_j = \begin{cases} 1 & \text{with probability } (w.p.)\ q_1 \\ 0 & w.p.\ q_2 \\ -1 & w.p.\ 1 - q_1 - q_2 \end{cases}$$

where $j = 1, 2, \ldots, k$. The tree model proceeds to construct a tree in the following way

1. Set $s_j = 1$ for $j = 1, ..., k$.

2. Choose a cutoff κ for between 0 and 1 for the branch probabilities.

3. Starting with s_1, consider the likelihood of $\boldsymbol{Y}$ by averaging over the different possibilities of s_j

$$f(\boldsymbol{Y}|q_1, q_2, \beta_0, \boldsymbol{\beta}_c) = q_1 L(\beta_0, \beta_{c,1}) + q_2 L(\beta_0, \beta_{c,0}) + (1 - q_1 - q_2) L(\beta_0, \beta_{c,-1}), \tag{2.4}$$

where for $h = 1, 0, -1$: $L(\beta_0, \beta_{c,h}) = \prod_{i=1}^{n} p_h^{Y_i} (1-p_h)^{Y_i}$ and $p_1 = \frac{\exp(\beta_0 + \beta_{c,1} \sum_{j=1}^{k} X_{ij})}{1+\exp(\beta_0 + \beta_{c,1} \sum_{j=1}^{k} X_{ij})}$, $p_0 = \frac{\exp(\beta_0 + \beta_{c,0} \sum_{j=2}^{k} X_{ij})}{1+\exp(\beta_0 + c \sum_{j=2}^{k} X_{ij})}$, and $p_{-1} = \frac{\exp(\beta_0 - \beta_{c,-1} X_{i1} + \beta_{c,-1} \sum_{j=2}^{k} X_{ij})}{1+\exp(\beta_0 - \beta_{c,-1} X_{i1} + \beta_{c,-1} \sum_{j=2}^{k} X_{ij})}$.

4. Maximize the likelihood in (2.4) with respect to q_1, q_2, β_0, $\beta_{c,1}$, $\beta_{c,0}$ and $\beta_{c,-1}$ and obtain the maximum likelihood estimates $\hat{q}_1$, $\hat{q}_2$, $\hat{\beta}_0$, $\hat{\beta}_{c,1}$, $\hat{\beta}_{c,0}$ and $\hat{\beta}_{c,-1}$ respectively.

5. Allow allocations with estimated path probabilities greater than $\max\{q_1, q_2, 1 - q_1 - q_2\} \times \kappa$.

6. For each allowed allocation of s_1,
   `for` $j = 2 : k$

7. Repeat Steps 3-5 to find possible values for $s_j,\ j = 2, \ldots, k$.

8. Finally, obtain a tree, each branch of which represents a possible allocation of $(s_1, ..., s_k)$.

9. Once again, the weight of each branch p_l is calculated by taking the product of $\hat{q}$s in successive branching steps.

Let there be finally m total branches, with overall weights $p_1, p_2, ..., p_m$ respectively. We propose a weighted score test using the above data adaptive model-averaging approach to test for the null hypothesis H_0 in Section 2.2.4. We will refer to this approach as Likelihood-based Model Branching (LiMB) method.

2.2.4 A weighted score test

We propose a weighted score test that computes score test statistics for testing $H_0 : \beta_c = 0$ based on the model selected in each branch and subsequently averages all the score test statistics with their corresponding weights $p_1, ..., p_m$ to compute the final weighted score test statistic. More precisely, let the model selected in the l-th branch be given by

$$\text{Logit}\Pr(Y_i = 1) = \beta_0 + \beta_c g_i^l,\ \ i = 1, ..., n;\ l = 1, ..., m, \tag{2.5}$$

where $g_i^l = \sum_{j=1}^{k} X_{ij} s_j^l$ with $(s_1^l, ..., s_k^l)$, the allocation vector $\boldsymbol{s}$ selected in the l-th branch. After some routine algebra, the score test statistic for testing $H_0 : \beta_c = 0$ corresponding to the l-th branch can be shown as

$$Sc_l = \frac{\left[\sum_{i=1}^{n}(Y_i - \bar{Y})(g_i^l - \bar{g}_l)\right]^2}{\bar{Y}(1-\bar{Y})\sum_{i=1}^{n}(g_i^l)^2}$$

The weighted score test denoted by $wscore$ is defined as

$$wscore = \sum_{l=1}^{m} Sc_l p_l$$

The distribution of this data adaptive score test under the null hypothesis is not known, so one needs to use a permutation test or other simulation-based approach to derive a p-value for this weighted score test statistic $wscore$.

2.2.5 Paring the branches

For either pathfinding scheme, branch weights, $p_1, p_2, \cdots, p_m$, are computed. Thus, we can further narrow the plausible models by choosing a cutoff, $q_{\max}$, such that selected branches have weights greater than or equal to $p_{\max} \cdot q_{\max}$. After paring our branches, we can recalculate the $wscore$ for the best branches.

The simulation section described next discusses the advantages and tradeoffs of the $wscore$ approach for each of the two different pathfinding approaches and their pared versions compared to model selection, collapsing, and random effects methods in terms of their power in a variety of simulation scenarios and a real data analysis.

2.3 Results

2.3.1 Simulation study

In this section, the performance of the proposed weighted score tests implementing the model-averaging schemes is compared to the model selection approaches, such as Seq-aSum and Seq-aSum-VS described in section 2.2.1. We also compare these approaches to Sum test (Pan, 2009) and SKAT (Wu et al., 2011) with a weighted linear kernel. For this purpose, we simulate data as described in Basu and Pan (2011). In particular, we simulate k RVs each with MAF = 0.005 and each common variant (CV) with MAF = 0.2. To simulate the datasets, we generate a latent vector $\boldsymbol{Z} = (Z_1, ..., Z_k)'$ from a multivariate normal distribution with a first-order auto-regressive (AR1) covariance structure with correlation $Corr(Z_i, Z_j) = \rho^{|i-j|}$ between any two latent components. For the purpose of this simulation, we have considered pairwise correlation of $\rho = 0$ and $\rho = 0.9$ which implies linkage equilibrium among the variants and strong linkage disequilibrium (LD) among the variants, respectively. Each component of the latent vector $\boldsymbol{Z}$ is then dichotomized to yield a haplotype, where the probability of $\boldsymbol{Z}$ being zero is the MAF corresponding to the RV. Next, we combine two independent haplotypes

and obtain genotype data $\boldsymbol{X}_i = (X_{i1}, ..., X_{ik})'$. The disease status Y_i is then generated from a logistic regression model with or without interaction. We have considered a sample of 500 cases and 500 controls.

We consider several simulation set-ups. We first simulate 10000 datasets under the null hypothesis of no association between the variants and the disease. For every set of RVs or mix of CVs and RVs, we estimate the null distribution of the test statistics based on 10000 replicates and determine the 95th percentile of the null distribution for each test statistic to use a critical value having $\alpha = 0.05$. We next compare the power of all the competing methods based on 10,000 simulated datasets for a variety of situations. When available, the asymptotic power is used by determining the number of times the calculated test p-value is less than 0.05. Otherwise, the empirical power is determined by the number of times the test statistic was greater than or equal to the 95th percentile determined from its null distribution.

We consider a variety of scenarios to test the performance of the proposed approaches. For demonstration, we first consider the situation displayed in Figure 2.1 where only the last RV is causal and the others are non-causal. Also, because Basu and Pan (2011) concluded model selection methods were a good compromise for a vast number of situations whereas the random effects and collapsing methods performance depended heavily on directionality of association, we consider the following scenarios: (1) four RVs are causal and (2) two RVs are causal while two RVs are protective. Both no LD and strong LD are used in these situations. Lastly, we consider cases when there is a mix of CVs and RVs to further understand the real data results. The order of the null and non-null variants are randomly assigned for each simulation. For each model-averaging method in Section 2.2.2 and Section 2.2.3, κ must be chosen so that the number of paths does not become too large. While many κ satisfy this, we present the results on model-averaging-based tests after fixing κ at 0.90 for LiMB and 0.95 and 0.99 for BUR. We also select $q_{\max} = 0.99$ for the pared version of our tests.

2.3.2 Simulation 1: Null distribution of the test statistic

Fig 2.2 shows the null density distribution of the test statistics of the BUR algorithm compared to Seq-aSum-VS while varying the number of null RVs included in the analysis. The average number of models averaged over is reported in the upper right corner.

The wscore and Seq-aSum-VS statistics increase as the number of RVs increase and have roughly the shape of a mixture of χ^2 distributions. However, due to the data adaptive procedure and the varying number of models averaged over, we cannot derive the theoretical null distribution. By construction, if the κ for the BUR algorithm was set to 1, the *wscore* statistic would be exactly equal to Seq-aSum-VS. We can see that when the mean number of branches is small in Fig 2.2, the distributions are almost equivalent. As the number of RVs increases, the dotted line representing BUR with $\kappa = 0.95$, which has the most models averaged over, begins to move to the left of the Seq-aSum-VS distribution. In other words, Seq-aSum-VS becomes stochastically greater than BUR with $\kappa = 0.95$. The BUR with $\kappa = 0.99$ remains very close to the Seq-aSum-VS statistic for all cases displayed because the average number of models averaged over remains low.

2.3.3 Comparing power among different approaches

Next we compare the power of different approaches for rare variant detection under different alternative models.

Simulation 2: Effect of order dependency Here, we demonstrate how model-averaging can address order dependency as in Fig 2.1. To do this, we set only the last RV to be causal with an odds ratio (OR) of 6. We then independently simulate 3, 7, 11, 15 and 19 null RVs in front of the causal RV. For each simulation setup we have generated 10000 replicates and have reported in Table 2.1 the empirical power of model selection methods (Seq-aSum, Seq-aSum-VS) and the direct model-averaging extension to Seq-aSum-VS, BUR, to demonstrate the reduction of order dependency by averaging over many models. The power is reported at a level of significance of 0.05. According to Table 2.1, as the number of null RVs before the one causal RV increases, the BUR ($\kappa = 0.95$) approach with pared branches becomes increasingly more powerful than Seq-aSum-VS. By exploring many paths and then excluding less likely paths, we reduce the dependency on the ordering of RVs and thus gain power to detect association here.

Simulation 3: Power comparison in the presence of no LD Here, we consider the situation where there is no LD between any two RVs, mimicking the situation where mutations are all completely random and independent of each other. Here we

Table 2.1: ($\boldsymbol{\alpha = 0.05}$) Demonstration of how model-averaging can reduce path dependency. In this disease model, only the last RV in order is causal with OR = 6. Empirical power listed in the table based on 10000 replicates with a number of non-causal RVs before the causal RV. There is no LD among the RVs.

No. of non-causal RVs	3	7	11	15	19
Seq-aSum	0.897	0.740	0.644	0.561	0.465
Seq-aSum-VS	0.920	0.811	0.703	0.615	0.527
BUR ($\kappa = 0.95$)					
Average no. branches	2.2	6.3	31.4	222	2130
Average no. pared branches	1.2	1.4	1.8	2.4	3.3
wscore	0.919	0.808	0.692	0.593	0.504
pared wscore	0.914	0.811	0.710	0.631	0.539
BUR ($\kappa = 0.99$)					
Average no. branches	1.3	1.5	2.1	3.3	5.4
Average no. pared branches	1.2	1.1	1.1	1.2	1.4
wscore	0.920	0.812	0.703	0.617	0.527
pared wscore	0.919	0.812	0.707	0.618	0.536

compare the power of our model-averaging approaches with several alternative methods. We first consider the situation where all 4 causal RVs share a common odds ratio (OR) of 2. We then simulate 0, 4, 8, and 12 null variants to study the impact of null variants on power. For each simulation setup we have generated 10000 replicates and reported the asymptotic or empirical power of a collapsing method (Sum), a random effect method (SKAT), model selection methods (Seq-aSum, Seq-aSum-VS), and model-averaging methods (BUR and LiMB) in Table 2.2.

As in Basu and Pan (2011), we see that the Sum test performs well above the other methods when there are very few non-causal variants present. As we increase the number of non-causal RVs, SKAT obtains an advantage over the Sum test. Model selection and model-averaging approaches obtain similar power to the collapsing approach as the number of non-causal RVs increases. Seq-aSum performs well when there is no null variant, but loses power compared to Seq-aSum-VS as in presence of null variants. The BUR ($\kappa = 0.99$) approach performs similarly to Seq-aSum-VS since only one or two paths are generally explored at $\kappa = 0.99$. The pared wscore for BUR ($\kappa = 0.95$) approach performs similarly to the BUR ($\kappa = 0.99$) approach whereas BUR ($\kappa = 0.95$)

loses little power due to averaging over too many null models. The LiMB approach does not perform well and has uniformly lower power than the other model averaging approaches.

Table 2.2: ($\boldsymbol{\alpha = 0.05}$) Power in table based on 10000 replicates for each situation with a number of non-causal RVs.

	OR = (2, 2, 2, 2)				OR = (4, 3, 1/3, 1/4)			
No. of non-causal RVs	0	4	8	12	0	4	8	12
Sum	0.710	0.482	0.362	0.289	0.501	0.315	0.237	0.191
SKAT	0.494	0.411	0.366	0.329	0.943	0.901	0.861	0.820
Seq-aSum	0.505	0.376	0.328	0.278	0.922	0.828	0.755	0.664
Seq-aSum-VS	0.500	0.397	0.337	0.285	0.906	0.836	0.768	0.681
LiMB ($\kappa = 0.90$)								
Average no. branches	1.5	2.2	3.8	6.9	1.1	1.7	2.7	4.5
Average no. pared branches	1.0	1.1	1.1	1.1	1.0	1.0	1.1	1.1
wscore	0.481	0.388	0.325	0.275	0.908	0.823	0.737	0.645
pared wscore	0.475	0.391	0.324	0.272	0.906	0.827	0.736	0.638
BUR ($\kappa = 0.95$)								
Average no. branches	1.4	3.4	16.4	97.0	1.5	4.2	18.5	121
Average no. pared branches	1.0	1.2	1.5	2.0	1.1	1.3	1.6	2.1
wscore	0.498	0.395	0.329	0.278	0.903	0.835	0.763	0.674
pared wscore	0.500	0.399	0.335	0.288	0.898	0.832	0.761	0.686
BUR ($\kappa = 0.99$)								
Average no. branches	1.1	1.4	1.9	2.8	1.1	1.4	1.9	2.9
Average no. pared branches	1.0	1.1	1.2	1.4	1.1	1.1	1.2	1.4
wscore	0.500	0.399	0.335	0.286	0.906	0.838	0.766	0.680
pared wscore	0.500	0.399	0.337	0.287	0.905	0.837	0.765	0.680

For the next scenario, the 4 causal RVs have various association strengths, OR = (4, 3, 1/3, 1/4). We then simulate 0, 4, 8, and 12 null variants to study the impact of null variants on power. Again for each simulation setup we have generated 10000 replicates and report the power of a collapsing method (Sum), a random effect method (SKAT), model selection methods (Seq-aSum, Seq-aSum-VS), and model-averaging methods (BUR and LiMB) in Table 2.2. As in Basu and Pan (2011), SKAT performs best for any given number of non-causal RVs, while the Sum test suffers dramatic power loss. Model selection is only slightly less powerful than the random effect

methods when there are few non-causal RVs. Once again, the BUR approach performs similarly to Seq-aSum-VS when only one or two paths are explored, and the LiMB approach does not perform well. Overall, model-averaging methods do not show much advantage over model selection methods when there is no LD among the variants.

Simulation 4: Power comparison in the presence of LD Table 2.3 considers the same cases as Table 2.2 except strong LD is present amongst the causal RVs and the non-causal RVs. When we have strong LD and causal RVs in the same direction, we can see that the collapsing and random effects methods perform similarly while model selection is lower powered. In this situation, there also appears to be no clear benefit to averaging over more models. This is because the non-causal RVs are strongly correlated with RVs that have effects in the same direction making them also have a marginal OR greater than 1. This means the best model is most likely one that is equivalent to the Sum test which is the first path that model selection and model-averaging consider. However, they are penalized by considering less likely models thereafter.

In the next situation, we have strong LD and causal RVs are associated in opposite directions. As in Basu and Pan (2011), model selection has a sizable advantage over collapsing and random effect methods when there are few non-causal RVs present. However, as non-causal RVs are introduced in the model, model selection loses its advantage to SKAT while still maintaining an advantage over the Sum test. The extension to model-averaging, though, appears to have better performance than model selection when a κ of 0.95 is used. Due to strong LD, the non-causal RVs will have similar effects to the causal ones. Now moving one such variant with negative directional effect to the '-1' category will be very similar to moving that variant to the '0' category, since the other correlated RVs will still be in the '+1' category, canceling the effect of this variant. Hence, an incorrect '0' allocation could be assigned to this variant. Until all of the variants with negative directional effect are moved to the '-1' category, we might not see much improvement in the likelihood. In this case, considering multiple allocations such as '-1' and '0' allocations for these variants would have better chance of finding the model that will significantly increase the likelihood of the data. Thus, BUR with $\kappa = 0.95$ presents an increase in power from Seq-aSum-VS by averaging over many models. We also can note that paring down to higher weighted models loses power in

Table 2.3: ($\boldsymbol{\alpha = 0.05}$) RV analysis when there is strong LD among the RVs. Power in table based on 10000 replicates for each situation with a number of non-causal RVs.

	OR = (2, 2, 2, 2)				OR = (4, 3, 1/3, 1/4)			
No. of non-causal RVs	0	4	8	12	0	4	8	12
Sum	0.999	0.971	0.904	0.838	0.263	0.287	0.296	0.284
SKAT	0.999	0.975	0.910	0.846	0.565	0.583	0.603	0.608
Seq-aSum	0.984	0.928	0.787	0.672	0.717	0.637	0.569	0.555
Seq-aSum-VS	0.984	0.925	0.799	0.668	0.710	0.641	0.577	0.552
LiMB ($\kappa = 0.90$)								
Average no. branches	1.2	2.3	4.8	8.7	1.2	2.3	4.1	6.8
Average no. pared branches	1.0	1.0	1.1	1.1	1.0	1.0	1.1	1.1
wscore	0.981	0.919	0.725	0.589	0.722	0.654	0.579	0.542
pared wscore	0.984	0.917	0.728	0.582	0.714	0.642	0.572	0.526
BUR ($\kappa = 0.95$)								
Average no. branches	3.4	33.9	485	10098	1.5	6.1	58.7	985
Average no. pared branches	1.1	1.2	1.4	1.8	1.0	1.1	1.4	1.9
wscore	0.984	0.916	0.775	0.653	0.709	0.648	0.594	0.578
pared wscore	0.983	0.902	0.720	0.573	0.708	0.643	0.569	0.560
BUR ($\kappa = 0.99$)								
Average no. branches	1.4	2.6	5.8	16.5	1.1	1.5	2.4	5.2
Average no. pared branches	1.1	1.1	1.2	1.4	1.0	1.1	1.2	1.3
wscore	0.984	0.919	0.788	0.667	0.710	0.642	0.577	0.554
pared wscore	0.984	0.919	0.790	0.669	0.710	0.640	0.579	0.556

the case of strong LD. Additionally, the unpared version of BUR with $\kappa = 0.95$ only has slightly less power than SKAT when there are 8 or 12 non-causal RVs present. This power comparison suggests that these model-averaging methods could be quite useful when the RVs are in strong LD with few causal variants in opposite directions of association and in the presence of few non-causal variants.

Simulation 5: Mix of CVs and RVs Here, we consider an analysis with a mixture of independent CVs and RVs. As recommended by Ionita-Laza et al. (2013), we have added SKAT-C which gives uniform weight to CVs rather than the Beta(MAF; 1, 25) weight of SKAT which severely downweights CVs. When CVs are mixed into a RV analysis, which is a usual scenario for scanning across a gene, the strength of contribution

of CVs and RVs greatly influences which method performs best. For Table 2.4, we first simulate 4 RVs with either shared common OR of 2 or two RVs with OR of 2 and the other two with OR of 1/2. We also simulated 3 moderately associated CVs of either OR = (1.2, 1.2, 1.2) or OR = (1.2, 1.2, 0.8). We then simulate 1, 5, 9, and 13 independent null CVs to study the impact of null variants on power. Because SKAT underweights these CVs, it suffers huge power loss compared to the other methods and as expected, SKAT-C performs much better than SKAT. SKAT would only perform well if mostly RVs contribute to disease risk (Ionita-Laza et al., 2013). SKAT-C is the top method when both CVs and RVs contribute to the risk but Seq-aSum-VS and BUR perform almost as well when there are a small number of null variants. Sum, as usual, suffers huge power loss in the presence of opposite directional effects. For the last situation, we make all of the RVs null and simulate 3 associated CVs with OR = (1.2, 1.2, 1.2). Unlike before, we can see that if only the CVs are associated, model selection and model-averaging performs well above the competitors, and BUR with $\kappa = 0.95$ shows improvement over Seq-aSum-VS when there are not many null CVs. For all of these situations, we see that Seq-aSum-VS and BUR do quite well especially if there are few null variants. This type of situation would be very common while scanning across a gene because variants in a window are likely to be in high LD and thus have a non-null effect.

2.3.4 Sanofi Data

Genomic intervals covering two genes that encode the endocannabinoid metabolic enzymes, FAAH and MGLL, were sequenced in 289 individuals of European ancestry using the Illumina GA sequencer (Bansal et al., 2011). Ancestry was determined using a panel of ancestry informative markers and individuals with an outlying genetic background were removed from the analysis. Sequencing was done using 36 base pair reads. The median coverage was 60X across the individuals sequenced. The program MAQ was used for alignment and variant calling, resulting in 1410 high quality single nucleotide variants (SNVs; 228 in the FAAH gene and 1182 in the MGLL gene) which were used for association analysis. The sequenced regions were captured using long range PCR and represented a total of 188,270 nucleotides. The 289 individuals included 147 normal controls (Body Mass Index (BMI) < 30) and 142 extremely obese cases (BMI > 40).

Each region was analyzed separately with a sliding window of 1000 bp in length. The size of this sliding window was chosen to ensure a reasonably small number of SNVs being analyzed at one time. The number of variants included in any window of either gene varies from 2-25 but about 90% included 5-15 SNVs. There were both common and rare (MAF ≤ 0.01) variants in the windows. Table 2.5 shows the distribution of the RVs and CVs in the reported windows. When available, we used the asymptotic distribution to calculate p-values. Otherwise, 1000 permutations were used to calculate p-values at each sliding window. Due to the poor performance of LiMB in the simulations, we have dropped it from the real data results. Because we have a mix of RVs and CVs, we have added SKAT-C which upweights CVs as compared to SKAT (Ionita-Laza et al., 2013). At each window, we recorded the minimum p-value of the competing methods. An additional 9000 permutations were performed for the most significant windows of each gene. To measure LD, we use the D' statistic (Lewontin, 1964). The mean LD in each of the sliding windows was moderate with $D' = 0.448$ and 0.449 in the FAAH and MGLL genes, respectively. D' ranged from 0.012 to 0.9996 for all sliding windows.

Figures 2.3 and 2.4 plot the $-\log_{10}$(p-values) of each window as it slides across FAAH and MGLL, respectively. The most significant windows of the 228 in the FAAH gene and the ten most significant windows of the 1182 in the MGLL gene are reported in Table 2.5 with bolded p-values for the best p-value in each window. We also denote the order of the sliding window and its starting genomic location.

Like in previous analyses (Bansal et al., 2011), the analysis shows little significant association of the FAAH gene with obesity in Fig 2.3. None of the p-values come close to the multiple comparisons level of significance of 4.06 (Bansal et al., 2011). We can see that the Sum test and model selection without variable selection drown out the faint signals shown by the other methods. The BUR approaches perform almost identically to Seq-aSum-VS.

The MGLL gene does show some suggestion of consistency of a signal in the right-most region in Fig 2.4. BUR with $\kappa = 0.95$ seems to amplify the signal in this area as compared to Seq-aSum-VS. Also, besides the middle-most region, SKAT appears to lack most of the signal shown by the other methods. Seq-aSum-VS and BUR seem to capture both the middle and right-most feature of the MGLL gene whereas the other methods only capture one or the other. SKAT-C also captures these regions, but with

the exception of the few windows shown in Table 2.5, Seq-aSum-VS and BUR show more significance. From Table 2.4, we might hypothesize that the right-most region has some moderate CV effects which SKAT fails to detect but SKAT-C, model selection, and model-averaging do detect. SKAT performs best when RVs contribute most to the risk such as in the windows of FAAH. Meanwhile, SKAT-C performs well when CVs and RVs are both contributing as they may be the case in the last few windows shown in Table 2.5. Model selection and BUR do quite well if CVs contribute to the risk, and from our simulations, they perform best when CVs contribute to most of the risk as they may in windows 1228 and 1229.

By looking at the top p-values in Table 2.5, we can assess the potential gains that model-averaging has over model selection. First of all, because the null distribution of model-averaging is stochastically smaller than model selection as we decrease κ, we can see that even when only one path is selected for BUR with $\kappa = 0.95$, it usually obtains a lower p-value than model selection and its close counterpart BUR with $\kappa = 0.99$. Also, out of the 16 top windows that BUR with $\kappa = 0.95$ has multiple paths averaged over, 10 of them produce a better p-value than when only one path is chosen by Seq-aSum-VS.

2.4 Discussion

In this chapter, we have studied the performance of several weighted score tests implementing model-averaging approaches and compared them to their competitors in detection of rare variants. It has been well documented (Basu and Pan, 2011) that no method is uniformly most powerful. Each method is very dependent on the underlying unknown true model. We have shown through simulation that each method has situations where it performs better than its competitors. However, through our simulations and the real data analysis we found that the Seq-aSum-VS and BUR approaches maintain reasonable power in almost all situations and never suffer huge power loss unlike the other methods, particularly when we have both CVs and RVs in the analysis and the CVs strongly contribute to disease risk. In fact, model selection and BUR were some of the top methods in all simulations when there were a low number of null variants. This situation would be very typical while scanning across a causal gene because variants in a window are likely to be in high LD with the causal variant and thus all have a non-null

effect. We have focused on the comparison of model-averaging with model selection approaches. While the advantages of model-averaging have been well documented in the prediction literature (Raftery et al., 1997), we studied the advantage of model-averaging over model selection when our purpose is inference. As shown in simulation studies and real data analysis, model-averaging over a limited number of models showed a power gain over model selection, but the power gain was not substantial in most simulation setups. One possible explanation could be that the model selection approach already implements a dimension reduction strategy which requires estimation of only three parameters for each model. Due to the small number of parameters, the uncertainty in model selection decreases and the advantage of model-averaging over model selection becomes less significant.

The model-averaging approach was proposed to reduce the dependency of the model selection approaches such as Seq-aSum and Seq-aSum-VS on the sequential selection of the SNPs. The performance of a model selection approach would depend on the order at which the SNPs were selected sequentially. A model-averaging approach, on the other hand, reduces this order dependency. In addition to this reduction of order dependency, we saw that averaging over more models can present a gain in power over one model, particularly when variants are in strong LD and when there is a mix of causal and protective RVs. We also saw in our simulations and possibly the real data analysis that BUR with $\kappa = 0.95$ had significant gains over model selection when CVs strongly contribute to the risk with only a small number of null variants. So while model selection was presented as a good middle approach for any alternative disease model in Basu and Pan (2011), model-averaging is perhaps more advantageous because it performs as well or better depending on the truth.

If covariates were to be added, permutation of the outcomes would no longer suffice if the goal is to test the genetic effect. Instead, one could fit a model with only the covariates and then use a parametric bootstrap using the estimated covariate effects to simulate the same number of datasets that you would use in a permutation test(Bůžková et al., 2011). We then proceed in a similar fashion as in a permutation approach, where we perform model-averaging on the simulated set and compare our test statistic to the bootstrapped test statistics.

In general, the BUR approach with 0.95 cutoff performed better than the BUR

approach with 0.99 cutoff, which indicates a clear benefit from averaging over more models since it accounts for the model uncertainty. When too many models are averaged over with independent RVs, the BUR approach with 0.95 cutoff is still better but we need to pare down the branches. On the other hand, one big limitation of model-averaging is the number of models averaged over. It became too computationally intensive once we considered more than 20 variants. Hence, model-averaging has an advantage over model selection when we consider a small to moderate number of variants. From the simulations, we would recommend using the BUR approach with $\kappa = 0.95$ in order to search a wide array of models. If the variants in the SNV-set are independent or weakly correlated, we would also recommend paring down to only the top models to reduce the number of models to average over. Use of this recommended application of our proposed model-averaging is illustrated in the real data section. A sliding window with 5-20 variants could give us optimal performance of the model-averaging approach. In the future, we intend to compare this model-averaging approach with a model-averaging approach with distinct parameters for each directional effect in the BUR approach, while undergoing variable selection. We believe there would be substantial power gain over the reduced model with same effect size for both directions.

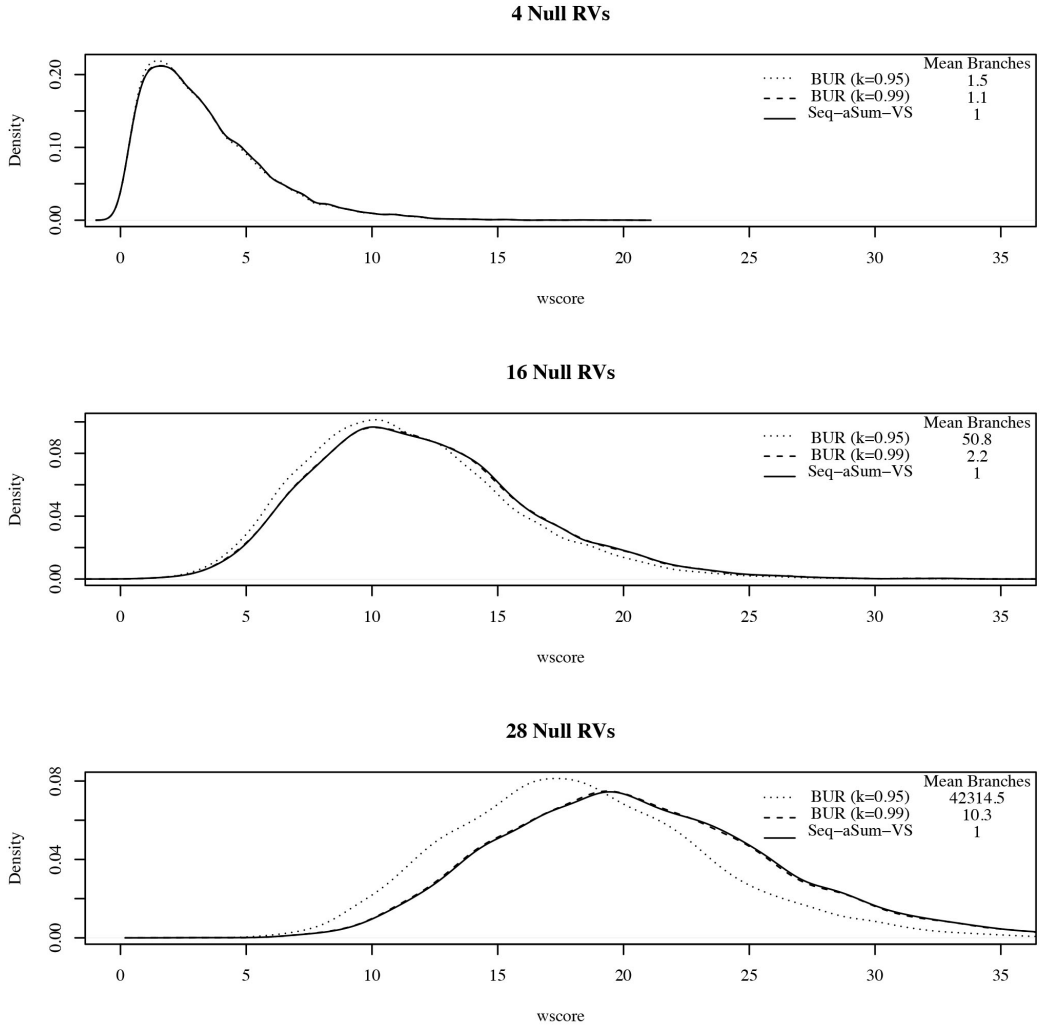

Figure 2.2: Density plot of model selection and BUR model-averaging approach test statistics under the null situation where all RVs are non-causal given a RV-set size of 4 (top), 16 (middle), or 28 (bottom). The x-axis is the value of the test statistic and the y-axis is the density of the distribution. Plotted in each figure are Seq-aSum-VS, BUR ($\kappa = 0.95$), and BUR ($\kappa = 0.99$) as solid, dotted, and dashed lines, respectively. The number of branches averaged over by these model-averaging approaches is in the upper right table of each plot.

Table 2.4: ($\boldsymbol{\alpha = 0.05}$) Analysis with a combination of RVs and CVs. Power in table based on 10000 replicates for each situation with a mix of common and rare variants. The CVs and RVs are all independent of each other.

RVs	OR = (2, 2, 2, 2)				OR = (2, 2, 1/2, 1/2)				OR = (1, 1, 1, 1)			
Associated CVs	OR = (1.2, 1.2, 1.2)				OR = (1.2, 1.2, 0.8)				OR = (1.2, 1.2, 1.2)			
Null CVs	1	5	9	13	1	5	9	13	1	5	9	13
Sum	0.845	0.576	0.428	0.327	0.141	0.098	0.082	0.072	0.430	0.321	0.255	0.207
SKAT	0.496	0.501	0.488	0.490	0.404	0.391	0.390	0.387	0.227	0.169	0.145	0.119
SKAT-C	0.791	0.717	0.674	0.631	0.778	0.684	0.633	0.590	0.494	0.422	0.378	0.327
Seq-aSum	0.729	0.463	0.357	0.292	0.669	0.436	0.341	0.292	0.651	0.446	0.368	0.311
Seq-aSum-VS	0.760	0.523	0.432	0.333	0.672	0.449	0.379	0.304	0.681	0.465	0.412	0.329
LiMB ($\kappa = 0.90$)												
Average no. branches	1.5	2.3	3.9	8.3	1.7	3.2	6.4	12.6	1.4	1.8	2.7	4.7
Average no. pared branches	1.0	1.1	1.1	1.2	1.1	1.2	1.2	1.2	1.0	1.0	1.0	1.0
wscore	0.233	0.174	0.069	0.049	0.531	0.300	0.167	0.127	0.144	0.129	0.110	0.061
pared wscore	0.230	0.171	0.073	0.053	0.526	0.297	0.174	0.133	0.147	0.130	0.112	0.063
BUR ($\kappa = 0.95$)												
Average no. branches	8.0	27.8	161	1245	7.2	25.7	137	965	3.6	13.8	78.2	577
Average no. pared branches	2.0	2.5	3.2	4.3	1.9	2.6	3.5	4.9	1.1	1.3	1.7	2.4
wscore	0.766	0.515	0.412	0.316	0.677	0.451	0.371	0.304	0.692	0.466	0.398	0.324
pared wscore	0.761	0.527	0.429	0.329	0.668	0.451	0.374	0.304	0.690	0.481	0.415	0.332
BUR ($\kappa = 0.99$)												
Average no. branches	1.7	2.3	3.5	5.8	1.8	2.5	3.7	6.0	1.3	1.8	2.7	4.4
Average no. pared branches	1.4	1.6	1.9	2.3	1.6	1.9	2.2	2.7	1.1	1.1	1.3	1.5
wscore	0.761	0.524	0.431	0.333	0.672	0.449	0.377	0.304	0.683	0.466	0.412	0.329
pared wscore	0.760	0.525	0.431	0.334	0.670	0.447	0.376	0.305	0.682	0.469	0.412	0.331

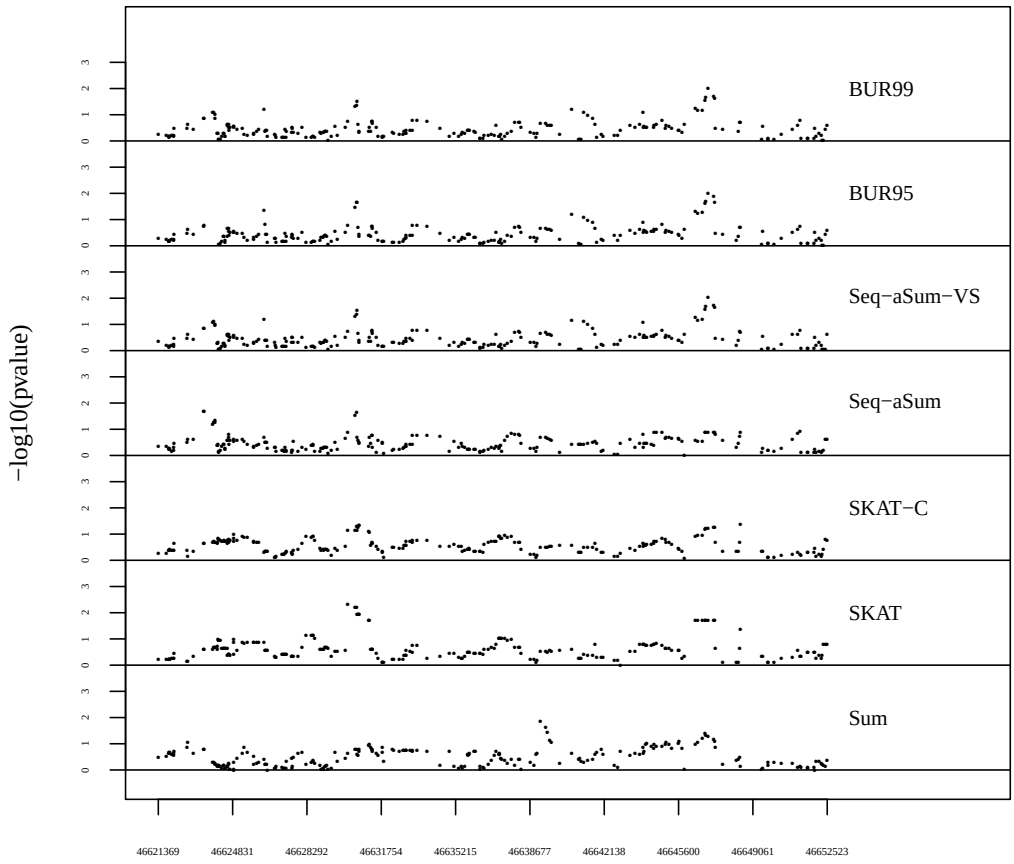

Figure 2.3: A sliding window analysis of FAAH gene for each method from top to bottom: BUR with $\kappa = 0.95$ (BUR95), BUR with $\kappa = 0.99$ (BUR99), Seq-aSum-VS, and Seq-aSum, SKAT-C, SKAT, and Sum. A window size of 1000 bp is used. The $-\log_{10}$(p-value) for each window is plotted on the y-axis. The beginning genomic location of each window is plotted across the x-axis. Each point represents the $-\log_{10}$(p-value) of one window.

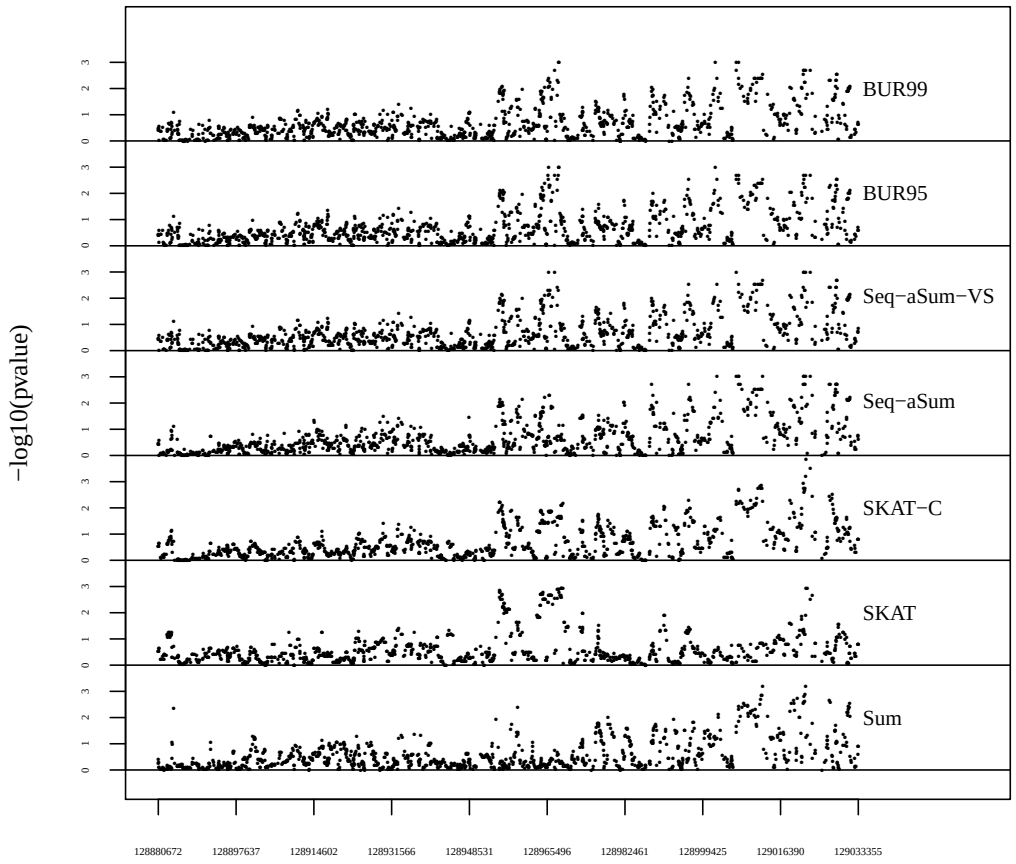

Figure 2.4: A sliding window analysis of MGLL gene for each method from top to bottom: BUR with $\kappa = 0.95$ (BUR95), BUR with $\kappa = 0.99$ (BUR99), Seq-aSum-VS, and Seq-aSum, SKAT-C, SKAT, and Sum. A window size of 1000 bp is used. The $-\log_{10}$(p-value) for each window is plotted on the y-axis. The beginning genomic location of each window is plotted across the x-axis. Each point represents the $-\log_{10}$(p-value) of one window.

Table 2.5: Top $-\log_{10}$(p-values) with window starting genomic location and the order of the frame for data analysis of both genes. BUR95 represents the BUR ($\kappa = 0.95$) approach and BUR99 represents the BUR ($\kappa = 0.99$) approach. The top p-value among these approaches is in bold for each window and the number of branches averaged over is listed in parentheses.

			Total					**Model selection**		**Model-averaging**	
Gene	Start Pos.	Window	RV	CV	Sum	SKAT	SKAT-C	Seq-aSum	Seq-aSum-VS	BUR95	BUR99
FAAH	46630186	85	3	6	0.647	**2.321**	1.159	0.872	0.738	0.756 (4)	0.736 (1)
	46630533	86	4	10	0.798	**2.213**	1.144	1.479	1.279	1.466 (2)	1.318 (2)
	46630611	87	4	10	0.787	**2.218**	1.302	1.503	1.372	1.611 (3)	1.326 (2)
	46630632	88	4	9	0.590	**1.952**	1.157	0.688	1.476	1.664 (1)	1.479 (1)
	46630673	89	3	9	0.561	**1.947**	1.260	0.580	0.647	0.705 (3)	0.624 (2)
	46630715	90	3	8	0.719	**1.947**	1.329	0.468	0.393	0.426 (9)	0.393 (1)
	46630718	91	3	7	0.788	**1.947**	1.347	0.500	0.304	0.355 (18)	0.317 (2)
	46639163	153	4	3	**1.854**	0.511	0.505	0.695	0.680	0.691 (1)	0.68 (1)
	46646969	199	4	2	1.265	1.709	1.230	0.859	1.880	**1.932 (1)**	1.88 (1)
	46647235	200	3	2	1.175	1.706	1.243	0.851	1.708	**1.785 (1)**	1.708 (1)
MGLL	128967903	966	8	3	0.583	**2.887**	1.626	0.730	2.387	2.796 (3)	2.432 (1)
	128967981	967	7	4	0.382	2.884	1.639	**3.301**	3.097	3.097 (2)	3.097 (1)
	128968059	968	6	5	0.342	2.785	1.662	1.146	2.620	**2.854 (1)**	2.62 (1)
	129002169	1228	0	5	1.165	0.556	1.107	2.538	3.000	**3.046 (2)**	3.000 (1)
	129002562	1229	1	5	1.205	0.561	1.195	2.658	2.482	**2.824 (6)**	2.469 (1)
	129006765	1251	5	2	1.438	0.104	2.245	**3.097**	2.959	2.569 (10)	2.959 (1)
	129007295	1255	2	2	2.317	0.750	**2.721**	2.678	2.658	2.569 (4)	2.658 (1)
	129012512	1277	3	2	**3.195**	0.839	2.262	2.699	2.553	2.523 (4)	2.553 (1)
	129021383	1324	5	4	2.562	1.261	**2.929**	2.699	2.398	2.399 (12)	2.524 (2)
	129021891	1329	1	3	3.191	1.882	**3.196**	2.854	2.602	2.569 (2)	2.602 (1)
	129021962	1330	1	3	1.245	2.913	**3.841**	1.903	1.572	1.582 (1)	1.572 (1)
	129022269	1331	0	3	1.338	2.914	**4.066**	2.056	1.791	1.81 (1)	1.791 (1)
	129022900	1332	2	2	0.988	2.503	**3.497**	3.000	2.959	2.959 (3)	2.959 (2)

Chapter 3

A combination test for detection of gene-environment interaction in cohort studies

3.1 Introduction

The interplay between genes and environments has been of much interest since the late 1800s when Galton famously introduced nature versus nurture (Galton, 1874). Today, we are better able to study this idea due to the advent of high-throughput technologies. While the focus over the past twenty years has been to identify single-nucleotide polymorphisms (SNPs) associated with complex diseases using genome-wide association studies (GWASs), the SNPs found in these studies explain very little disease heritability. There are several hypotheses to explain this "missing heritability" (Manolio et al., 2009), and one such hypothesis is the existence of gene-environment (GxE) interaction. Many examples of GxE interaction have been found in the past (Hunter, 2005). Identifying G-E interactions could potentially identify new variants associated with a disease and allow us to better understand its etiology.

In genome-wide association studies, investigators see to identify GxE interactions by scanning over all of the SNPs in the genome and testing for interaction between each SNP and an environment (Fan et al., 2016). While there are many ways to test

for a single interaction in the genome-wide approach (Mukherjee et al., 2011; Ko et al., 2013) researchers have had little success in finding any significant interactions (Hutter et al., 2013). One reason is the lack of statistical power for finding these interactions. In general, detection of an interaction requires four times the sample size than for detecting a main effect with comparable effect size (Thomas, 2010a). Also, testing each interaction individually suffers from issues similar to the single-SNP analyses in a GWAS. These tests lose power due to the correction for a large number of multiple comparisons and not modeling the likely joint effects between the SNPs, environments, and their interactions (Lesnick et al., 2007; Peng et al., 2009).

Because many studies lack the power to test for GxE interactions across the entire genome, many two-step strategies have been proposed to filter out SNPs that are not likely to have an interaction, and thus these studies gain power to detect interaction due to the reduction in the number of tests (Murcray et al., 2009, 2011; Dai et al., 2012; Hsu et al., 2012; Gauderman et al., 2013). One strategy is to select a group of SNPs that meet an arbitrary p-value threshold in a GWAS and test for pairwise interactions between these SNPs and an environment (Thomas, 2010a; Kooperberg and Leblanc, 2008). While these two-step strategies reduce the multiple testing burden, they still fail to take advantage of the dependence among the variants due to linkage disequilibrium (LD) and thus may lose power.

Another alternative approach is to focus on potential variants within a candidate gene for a disease to look for GxE interaction. This approach has demonstrated effectiveness as several studies have reported replication of the initial findings of GxE interaction (Simonds et al., 2016). Researchers have recently begun grouping SNPs into candidate genes or biological pathways to reduce the number of interactions tested and possibly increase power by pooling the effects of multiple genetic variants. However, joint modeling of the SNPs and environment quickly becomes intractable as the number of SNPs increases. Alternatively, several new gene-based tests for GxE interaction can efficiently deal with a large number of SNPs within a gene (Pan et al., 2011; Yu et al., 2012; Jiao et al., 2013; Lin et al., 2013, 2015; Wang et al., 2015). However, tests that use a gene-based summary measure as shown by Lin et al. (2015) can be biased and have inflated type I error especially if the genes and environment are not independent.

In this chapter, we consider testing for GxE interaction between a set of SNPs

from a candidate gene and an environmental factor in unrelated individuals. We extend a recently proposed method of combining the score statistic for genetic main effects (Pan et al., 2014) to the interaction testing problem. We also extend the sequential score-based test proposed in Basu and Pan (2011), and propose an approximate and a resampling-based test to detect GxE interaction. We study the performance of these tests under both the senarios of G-E independence and G-E dependence. Under realistic scenarios of G-E dependence, we demonstrate that tests based on a summary measure can maintain type I error and provide a powerful alternative to the other methods.

Finally, we have studied the performance of the methods using the Minnesota Center for Twin and Family Reseach (MCTFR) dataset. The MCTFR recently showed how particular environmental factors relate to substance abuse risk and interact with genetic risk (Hicks et al., 2009). We perform a gene-based GxE interaction analysis on the parent cohort of the MCTFR to study how genes interact with family climate to impact alcohol consumption.

3.2 Methods

Assume we have n unrelated individuals with Q common genetic variants from a candidate gene, K measured covariates, and an environmental factor. Let Y_i, E_i, $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{iK})$ be the phenotype, environmental factor, and K covariates for the i^{th} individual, respectively. Let $\mathbf{G}_i = (G_{i1}, \cdots, G_{iQ})$ be the minor allele counts for the Q variants, each standardized by its mean and standard deviation. Define $\mathbf{S}_i = (G_{i1}E_i, \cdots, G_{iQ}E_i) = (S_{i1}, \cdots, S_{iQ})$ to be the Q pairwise G-E interactions for the i^{th} individual. We will assume that our phenotype is continuous although all of these methods can be extended to binary phenotypes.

The SNPs, environments, and their pairwise interactions can be jointly modeled for the i^{th} individual by

$$Y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_{1,k} X_{ik} + \alpha_2 E_i + \sum_{q=1}^{Q} \alpha_{3,q} G_{iq} + \sum_{q=1}^{Q} \beta_q S_{iq} + \epsilon_i; \quad i = 1, ..., n \tag{3.1}$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2)$ and α_0, $\boldsymbol{\alpha}_1 = (\alpha_{1,1}, \cdots, \alpha_{1,K})$, α_2, $\boldsymbol{\alpha}_3 = (\alpha_{3,1}, \cdots, \alpha_{3,Q})$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_Q)$ are the effect estimates for the intercept, covariates, environment, SNPs,

and interactions, respectively. Here, we are interested in testing the null hypothesis that there is no G-E interaction $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Without using a dimension reduction approach, a likelihood ratio test (LRT) of this null hypothesis would require fitting a linear model with $2Q + K + 3$ parameters. With the size of the parameter space, the estimation procedures can quickly become intractable. If we are able to fit such a model, a test of the null hypothesis of no G-E interaction could lose power due to the large Q degrees-of-freedom (df) needed for the test. In the following subsections, we describe more powerful alternatives to the LRT.

3.2.1 MinP test

A simple approach to analyze whether there is interaction between any of the SNPs and an environmental factor is to model one interaction at a time. This reformulates Model 3.1 into Q separate models testing for the presence of a pairwise interaction between the environment and q^{th} SNP by

$$Y_i = \alpha_0^* + \sum_{k=1}^{K} \alpha_{1,k}^* X_{ik} + \alpha_2^* E_i + \alpha_{3,q}^* G_{iq} + \beta_q^* (E_i \times G_{iq}) + \epsilon_i^*; \quad i = 1, ..., n. \tag{3.2}$$

where α_0^*, $\boldsymbol{\alpha}_1^* = (\alpha_{1,1}^*, \cdots, \alpha_{1,K}^*)$, α_2^*, $\boldsymbol{\alpha}_3^* = (\alpha_{3,1}^*, \cdots, \alpha_{3,Q}^*)$ and $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_Q^*)$ are the effect estimates for the intercept, covariates, environment, SNPs, and interactions, respectively. To evaluate whether there is G-E interaction across the candidate region ($H_0 : \beta_q^* = 0$ for all q), we calculate the p-value for each G-E interaction separately and find the minimum p-value among the tests. To adjust for multiple testing, we use a Bonferroni correction by multiplying our minimum p-value obtained from the minP test by the number of G-E interactions tested.

This single-marker model, though computationally efficient, assumes a misspecified model for each genetic marker. Lin et al. (2013) showed that the asymptotic limits of the MLEs $(\hat{\alpha}_0^*, \hat{\boldsymbol{\alpha}}_1^*, \hat{\alpha}_2^*, \hat{\alpha}_{3,q}^*, \hat{\beta}_q^*)$ are generally not equal to the true values of $(\alpha_0, \boldsymbol{\alpha}_1, \alpha_2, \alpha_{3,q}, \beta_q)$ even when there is G-E independence. If the true disease model is a multi-marker model where multiple SNPs influence disease such as in Model 3.1, this minP test approach may lose power to detect interactions because it fails to model the joint effects of the SNPs, environments, and their pairwise interactions. If there is G-E dependence, this test may also have inflated type I error (Lin et al., 2013). We

investigate this in Section 3.3.

3.2.2 Score tests of interaction

Use of the score vector corresponding to the interaction parameters in Equation 3.1 can be efficient because the score vector is estimated under the null hypothesis for Model 3.1 which allows us to reduce the number of parameters that we explicitly need to estimate (Pan et al., 2011). Pan et al. (2014) proposed an adaptive sum of powered score tests (aSPU) to test for genetic main effects. Here, we extend the test to the G-E interaction framework by considering the score vector of the interactions. The G-E interaction score statistic for the q^{th} interaction of Model 3.1 can easily be derived as

$$U_{iq} = \frac{S_{iq}(Y_i - \boldsymbol{\mu}(\hat{\boldsymbol{\alpha}}))}{\hat{\sigma}_e^2}; \qquad i = 1, \cdots, n,\ q = 1, \cdots, Q$$

where $\boldsymbol{\mu}(\hat{\boldsymbol{\alpha}}) = \hat{\alpha}_0 + \sum_{k=1}^{K} X_{ik}\hat{\alpha}_{1,k} + E_i\hat{\alpha}_2 + \sum_{q=1}^{Q} G_{iq}\hat{\alpha}_{3,q}$ and $\hat{\alpha}_0$, $\hat{\boldsymbol{\alpha}}_1$, $\hat{\alpha}_2$, $\hat{\boldsymbol{\alpha}}_3$ and $\hat{\sigma}_e^2$ are estimated under the null model. The null model may be difficult to estimate if Q is large, so a ridge penalty on the genetic main effect $\boldsymbol{\alpha}_3$ can be used to improve estimation (Lin et al., 2013). The score-based family of SPU tests (Pan et al., 2014) is defined as

$$T_{SPU(\gamma)} = \sum_{q=1}^{Q} U_q^{\gamma}; \qquad \text{for a set of integers } \gamma \geq 1$$

where $U_q = \sum_{i=1}^{n} U_{iq}$. Under H_0, we assume that the asymptotic null distribution of the score vector $\mathbf{U} = (U_1, \cdots, U_Q) \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ holds where $\mathbf{V} = \sum_{i=1}^{n} \mathbf{U}_i\mathbf{U}_i^T$ and $\mathbf{U}_i = (U_{i1}, \cdots, U_{iQ})$. Therefore, we can generate B copies of the null score vector for which we calculate B copies of the SPU test $T_{SPU(\gamma)}^{(b)}$ where $b = 1, \cdots, B$. The p-value for a given γ is thus $P_{SPU(\gamma)} = (\sum_{b=1}^{B} I(|T_{SPU(\gamma)}^{(b)}| > |T_{SPU(\gamma)}|) + 1)/(B+1)$. Using the p-values for a set of γs, Pan et al. (2014) also proposes the aSPU test: $T_{aSPU} = \min_{\gamma\in\Gamma} P_{SPU(\gamma)}$. Using the same B copies of the null score vector, we can calculate the aSPU test statistic for each null score vector $T_{aSPU}^{(b)}$ and find the proportion of null aSPU test statistics that are smaller than our observed aSPU test statistic. Thus, $P_{aSPU} = (\sum_{b=1}^{B} I(T_{aSPU} > T_{aSPU}^{(b)}) + 1)/(B+1)$.

The Sequence Kernel Association Test (SKAT) (Wu et al., 2011) has been a popular method in this category for testing for association between a group of rare variants

(RVs) or SNPs and disease. Lin et al. (2013) and Lin et al. (2015) have extended SKAT to the Gene-Environment Set Assocation Test (GESAT) and interaction SKAT (iSKAT) which tests for interaction between an environment and a group of SNPs or RVs, respectively. In this paper we only consider common variants. So for Model 3.1, GESAT redefines $\beta_q \overset{\text{iid}}{\sim} N(0, \tau)$. By testing the null hypothesis $H_0 : \tau = 0$, GESAT equivalently tests $H_0 : \boldsymbol{\beta} = \mathbf{0}$. The score test of τ leads to the GESAT test statistic of $Q = (\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\alpha}}))^T \mathbf{S}\mathbf{S}^T (\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{\alpha}}))$ where $\boldsymbol{Y} = (Y_i, \cdots, Y_n)^T$, $\mathbf{S} = (\mathbf{S}_i, \cdots, \mathbf{S}_n)^T$, and $\boldsymbol{\mu}(\hat{\boldsymbol{\alpha}})$ contains the fitted values estimated under the null model of Model 3.1. To avoid problems with estimation of the genetic main effect, Lin et al. (2013) implement a ridge penalty on the genetic main effect $\boldsymbol{\alpha}_3$. Q asymptotically follows a mixture of chi-squares distribution and a p-value can be obtained using characteristic function inversion (Davies, 1980; Lin et al., 2013). Note that the GESAT test statistic without a ridge penalty is equivalent to the SPU test statistic with $\gamma = 2$.

3.2.3 G-E interaction using summary measures

Summary Measures based on Main Effect Model

An approach to reduce the dimension of the interaction test based on Model 3.1 is to use some summary score for the gene instead of modeling the main effects of inividual SNPs within the gene. The Sum test or Burden test is a popular choice in this category (Pan, 2009; Lin et al., 2015). Instead of using the Model in Equation 3.1, we use the following model:

$$Y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_{1,k} X_{ik} + \alpha_2 E_i + \alpha_G \sum_{q=1}^{Q} G_{i,q} + \beta^* \sum_{q=1}^{Q} G_{i,q} \times E_i + \epsilon_i; \quad i = 1, ..., n. \tag{3.3}$$

We assume same effect size and direction for the all the SNPs within a gene and test for the null hypothesis $\beta^* = 0$. The test statistic follows χ^2_1 distribution under the null hypothesis.

One could also use the likelihood-based data adaptive scheme as discussed in Basu and Pan (2011). Basu and Pan (2011) use a data adaptive scheme to create a gene-based summary measure to test whether that score is associated with disease. The test produced by this scheme is referred to as the sequential adaptive sum (Seq-aSum-VS)

test. While the Seq-aSum-VS approach is focused toward detection of main effects of SNPs, RVs, or a combination, it can be extended to test for G-E interaction. Once the summary measure is selected, once could test for interaction between the summary score and the environment.

We first consider testing for G-E interaction using the gene-based summary measure constructed from a main-effect-only model by Seq-aSum-VS. To do this, we first consider the following main-effect model:

$$Y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_{1,k} X_{ik} + \alpha_2 E_i + \alpha_G \sum_{q=1}^{Q} \gamma_q G_{iq} + \epsilon_i; \quad i = 1, ..., n \tag{3.4}$$

where $\sum_{q=1}^{Q} \gamma_q G_{iq}$ is the gene-based summary for the i^{th} individual, $\gamma_q = w_q s_q$ is a weight assigned to variant q, $s_q = 1$ or -1 indicating whether the effect of variant q is positive or negative, $s_q = 0$ indicating the exclusion of variant q from the model, and α_G is the effect estimate for this combination. We generally take w_q=1 and choose γ_q using the data adaptive scheme from Basu and Pan (2011). To find the optimal selection of γ_qs we proceed through the following steps

1. Initialize $\gamma_1 = \cdots = \gamma_Q = 1$
2. `for` j in $1 : Q$
    - Find the maximized likelihood corresponding to $\gamma_j = -1, 0, 1$ of Model 3.4
    - Set γ_j to be the value that corresponds to the largest maximized likelihood among the three.

Once the optimal allocation $\hat{\gamma}_q$ for $q = 1, \ldots, Q$ is selected for all the variants, we test for interaction between the gene-based summary measure generated by the optimal allocation $\hat{\boldsymbol{\gamma}}$ and genetic data G with the environment E based on the following model:

$$Y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_{1,k} X_{ik} + \alpha_2 E_i + \alpha_G \sum_{q=1}^{Q} \hat{\gamma}_q G_{iq} + \beta^* \left(E_i \times \sum_{q=1}^{Q} \hat{\gamma}_q G_{iq} \right) + \epsilon_i'; \; i = 1, ..., n. \tag{3.5}$$

With this setup, we are interested in testing only one parameter ($H_0 : \beta^* = 0$) rather than q parameters in Model 3.1. To test our null hypothesis, we obtain the maximized

likelihood under Model 3.5 and the null model with $\beta^* = 0$ and calculate the likelihood ratio test statistic for β^*. This null distribution of this test corresponds to a chi-square distribution with one df. We refer to this method as interaction testing using Seq-aSum for a Gene (iSeq-aSum-G). This test would have a power advantage over the interaction tests previously discussed if the main effect scoring captures the non-null variants involved in the interaction as well as the direction of the interaction.

This is similar to designing an interaction test based on the burden summary measure for the genetic main effect as discussed above. We obtain the burden test by setting $\gamma_q = 1$ for all q in Equation 3.4. However, our approach may have better power than the burden approach since the former applies model selection in the final choice of the scoring. One limitation of this test is that it may only be valid under the assumption of G-E independence or under the assumption that the associated SNPs have the same effect sizes. Note that in a gene with highly correlated SNPs, it is not unlikely for the associated SNPs to have similar effect sizes.

One issue with the tests in this category that they may not maintain the correct type I error rate in the presence of G-E dependence. Moreover, as illustrated in Lin et al. (2015), even under G-E independence, the homoscedasticity assumption may still be violated, which could impact the type I error of this test under G-E independence. Nevertheless, one should always consider the possibility of gain in power by using these summary measure-based tests due to substantially fewer degrees-of-freedom in the interaction.

Summary Measures based on Interaction Effect Model

Testing for interaction between a gene-based summary measure and an environment can be very restrictive even under G-E independence and may suffer from loss of power when a main effect model cannot capture interactions. Instead, we can extend Seq-aSum-VS for interaction testing by incorporating the sequential algorithm on the interaction itself rather than the main effects. We do this by using an interaction summary measure in the following model

$$Y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_{1,k} X_{ik} + \alpha_2 E_i + \sum_{q=1}^{Q} \alpha_{3,q} G_{iq} + \beta_c \sum_{q=1}^{Q} \gamma_q S_{iq} + \epsilon_i; \quad i = 1, ..., n. \tag{3.6}$$

where $\sum_{q=1}^{Q} \gamma_q S_{iq}$ is the interaction summary measure for the i^{th} individual, γ_q is the direction of association of the interaction between the q^{th} SNP and the environment. We use a data adaptive scheme to choose our optimal allocation for $(\gamma_1, \cdots, \gamma_Q)$. We find the optimal allocation of the interactions using the following sequential algorithm:

1. Initialize $\gamma_1 = \cdots = \gamma_Q = 1$
2. `for` j in $1 : Q$
   - Set $\gamma_j = -1$, 0, or 1 and calculate the likelihood ratio (LR) corresponding to a test of the null hypothesis $H_0 : \beta_c = 0$ in Model 3.6
   $$\mathrm{LR}(\beta_c, \gamma_j)) = -2 \log \left(\frac{\max_{\theta, \beta_c} \Pr(Y|X, E, G, S, \gamma_j, \beta_c)}{\max_{\theta, \beta_c = 0} \Pr(Y|X, E, G)} \right) \tag{3.7}$$
   where θ contains the main effect and variance parameters and γ_j is the current allocation of j-th locus with $\gamma_j = -1$, 0, or 1.
   - Set γ_j to be the value that corresponds to $\max_{\gamma_j \in \{-1,0,1\}} \mathrm{LR}(\beta_c, \gamma_j)$

Note that under the null hypothesis, $\Pr(Y|X, E, G)$ does not depend on the allocation. We use the maximum likelihood to estimate all of the parameters. This sequential algorithm avoids searching all 3^Q possible allocations and instead searches through $3Q + 1$ allocations. The final test statistic T_I is computed by taking the maximum LR of the $2Q + 1$ allocations browsed.

We will denote this approach as iSeq-aSum for Interaction (iSeq-aSum-I). For Q independent variants, the test statistic T_I will have a chi-square distribution with approximately $3Q/4$ df for large Q (See Appendix A.1 for proof). However, if there is correlation among the variants, the p-values calculated by using this null distribution can be conservative. Thus, in these situations, we propose using a parametric bootstrap as suggested by Bůžková et al. (2011) to calculate p-values for iSeq-aSum-I. To produce parametric bootstrap samples, we estimate $\hat{\alpha}_0$, $\hat{\boldsymbol{\alpha}}_1$, $\hat{\alpha}_2$, $\hat{\boldsymbol{\alpha}}_3$, and $\hat{\sigma}_e^2$ by fitting the main effects only model

$$Y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_{1,k} X_{ik} + \alpha_2 E_i + \sum_{q=1}^{Q} \alpha_{3,q} G_{iq} + \epsilon_i; \quad i = 1, ..., n \tag{3.8}$$

and obtain $\boldsymbol{\mu}(\hat{\boldsymbol{\alpha}})$. We then simulate B parametric bootstrap samples by $\boldsymbol{Y}^{(1)}, \cdots, \boldsymbol{Y}^{(B)} \sim \mathcal{N}\left(\boldsymbol{\mu}(\hat{\boldsymbol{\alpha}}), \hat{\sigma}_e^2 \mathbf{I}\right)$. We propose two strategies using the parametric bootstrap.

Bootstrap Strategy 1 The first strategy is to use a full parametric bootstrap. To do this, we obtain B null iSeq-aSum-I test statistics $T_I^{(b)}$ where B is 1000 or larger. We calculate the parametric bootstrap p-value as $(\sum_{b=1}^{B}(T_I > T_I^{(b)}) + 1)/(B + 1)$. However, this process can be very computationally intensive.

Bootstrap Strategy 2 To avoid the high computational cost of a full parametric bootstrap, our second strategy uses the parametric bootstrap to approximate the df for our chi-square distribution in the presence of correlated SNPs. Using only 100 parametric bootstrap samples, we calculate a null iSeq-aSum-I test statistic for each sample and estimate the mean of the null distribution. To calculate p-values for this strategy, the mean is used as the estimate the df for the chi-square distribution instead of $3Q/4$.

Combining our scoring approaches

The power of the above approaches will greatly depend on the underlying true interaction model. If the non-null interaction effects are captured by a main effect model, iSeq-aSum-G may outperform iSeq-aSum-I because it is using fewer degrees of freedom for interaction testing, especially if the effect sizes are not different among the variants within the gene under consideration. However, in the absence of main effects, the performance of iSeq-aSum-G will likely suffer while iSeq-aSum-I would be more powerful. Here, we propose a unified scoring approach to take advantage of the strengths of each of the scoring methods. We define the iSeq-aSum-min test statistic as $T_{\min} = \min\{p_G, p_I\}$ where p_G and p_I are the p-values for iSeq-aSum-G and iSeq-aSum-I, respectively. If there is G-E dependence, iSeq-aSum-G can have inflated type I error. However, by combining iSeq-aSum-G with iSeq-aSum-I, which is unaffected by G-E dependence, iSeq-aSum-min reduces the impact of G-E dependence on type I error. We use either of the two parametric bootstrap strategies to calculate the p-value for our combination test as described below.

Bootstrap Strategy 1 For the full parametric bootstrap strategy, we use the same bootstrap samples ($B \geq 1000$) and calculate the p-values for iSeq-aSum-G and iSeq-aSum-I for the b^{th} parametric bootstrap sample and obtain the minimum $T_{\min}^{(b)}$. The p-value of iSeq-aSum-min can then be calculated as $\left(\sum_{b=1}^{B} I\left\{T_{\min} > T_{\min}^{(b)}\right\} + 1\right)/(B+1)$.

Bootstrap Strategy 2 Alternatively, if we calculate the p-value of iSeq-aSum-I using Bootstrap Strategy 2, we use a simple Bonferroni Correction to calculate the p-value of iSeq-aSum-min. While this correction may be conservative, we avoid the additional computational cost of Bootstrap Strategy 1.

3.3 Results

For our simulations, we used the genotype data of the 3202 parents of the MCTFR to simulate a phenotype. We have selected two candidate genes for alcoholism (Olfson and Bierut, 2012), ADH1B and ALDH1A1, which have 11 or 50 SNPs genotyped by the Illumina 660W Quad array, respectively, and LD patterns as shown in Figure 3.1. For each gene, we considered n = 1000 and generated datasets by sampling genotypes without replacement from the parent population. We selected four causal SNPs for each gene from two different LD blocks as shown in Figure 3.1. Using these selected SNPs, we generated a dependent environment using a model similar to the one used in Lin et al. (2013):

$$E_i = \phi \sum_{k=1}^{4} G_{ik} + \eta_i \tag{3.9}$$

where G_{ik} for $k = 1, \cdots, 4$ is the set of the four selected SNPs and $\eta_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. If $\phi = 0$, there is G-E independence. We simulated each subject's alcohol phenotype from the linear model:

$$Y_i = \alpha_0 - 0.03\, X_i + 0.5\, E_i + \sum_{k=1}^{4} \alpha_3 G_{ik} + \sum_{k=1}^{2} \beta_1 S_{ik} + \sum_{k=3}^{4} \beta_2 S_{ik} + \epsilon_i; \quad i = 1, ..., n, \tag{3.10}$$

where $\alpha_0 = 0$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. A covariate X was simulated from a normal distribution with mean 50 and standard deviation 7 to mimic the age of the MCTFR parent

population. For our simulations, we assumed that all four selected SNPs interact with the environment.

We generated 10,000 datasets and 1,000 datasets to estimate the type I error and empirical power, respectively, at an $\alpha = 0.005$ level, which is the α-level used in Section 3.4. Using these datasets, we compared the performances of the minP test, aSPU, GESAT, burden, iSeq-aSum-G, iSeq-aSum-I, and iSeq-aSum-min using three simulations. For iSeq-aSum-I and iSeq-aSum-min, we used Bootstrap Strategy 2 from Section 3.2.3 to calculate p-values. Compared to the full parametric bootstrap (Bootstrap Strategy 1), Bootstrap Strategy 2 was faster with only a minimal power loss.

In Simulation 1, we evaluated the effect of G-E dependence on type I error. We assumed G-E independence for other simulations. In Simulation 2, we evaluated the power for detecting G-E interaction using ADH1B, a gene with only 11 SNPs and low LD, and ALDH1A1, a gene with 50 SNPs and high LD. In Simulation 3, we assessed how large the genetic main effect needs to be in order for iSeq-aSum-G to capture the interaction because iSeq-aSum-G relies on modeling interactions through the main effect model.

Simulation 1 We first compared the type I error of the different methods. To evaluate type I error, we set $\beta_1 = \beta_2 = 0$ in Equation 3.10. We considered both G-E independence and G-E dependence using the ADH1B gene to study the impact on type I error. The minP test and the summary measure-based tests can have inflated type I error if there is G-E dependence. To create G-E dependence, we varied ϕ in Equation 3.9 from 0 to 1. Using the genotypes of the parent population we sampled from, we estimated the correlation between the burden summary measure and environment displayed in Figure 3.2 by the formula derived in Appendix A.2. We set $\alpha_3 = 0.2$ or 0.5 so that the causal variants of ADH1B collectively explained 2% or 12%, respectively, of the total variation in our simulated phenotype.

The empirical Type I error rates of the methods are shown in Figure 3.2. When $\alpha_3 = 0.2$, all of the methods maintained correct type I error over the entire range of G-E correlation. However, if $\alpha_3 = 0.5$ and thus the causal SNPs explain a large proportion of variance, our simulations showed that the type I error became inflated for the minP test, iSeq-aSum-G, and the burden test for a G-E correlation greater than 0.07, 0.11, and

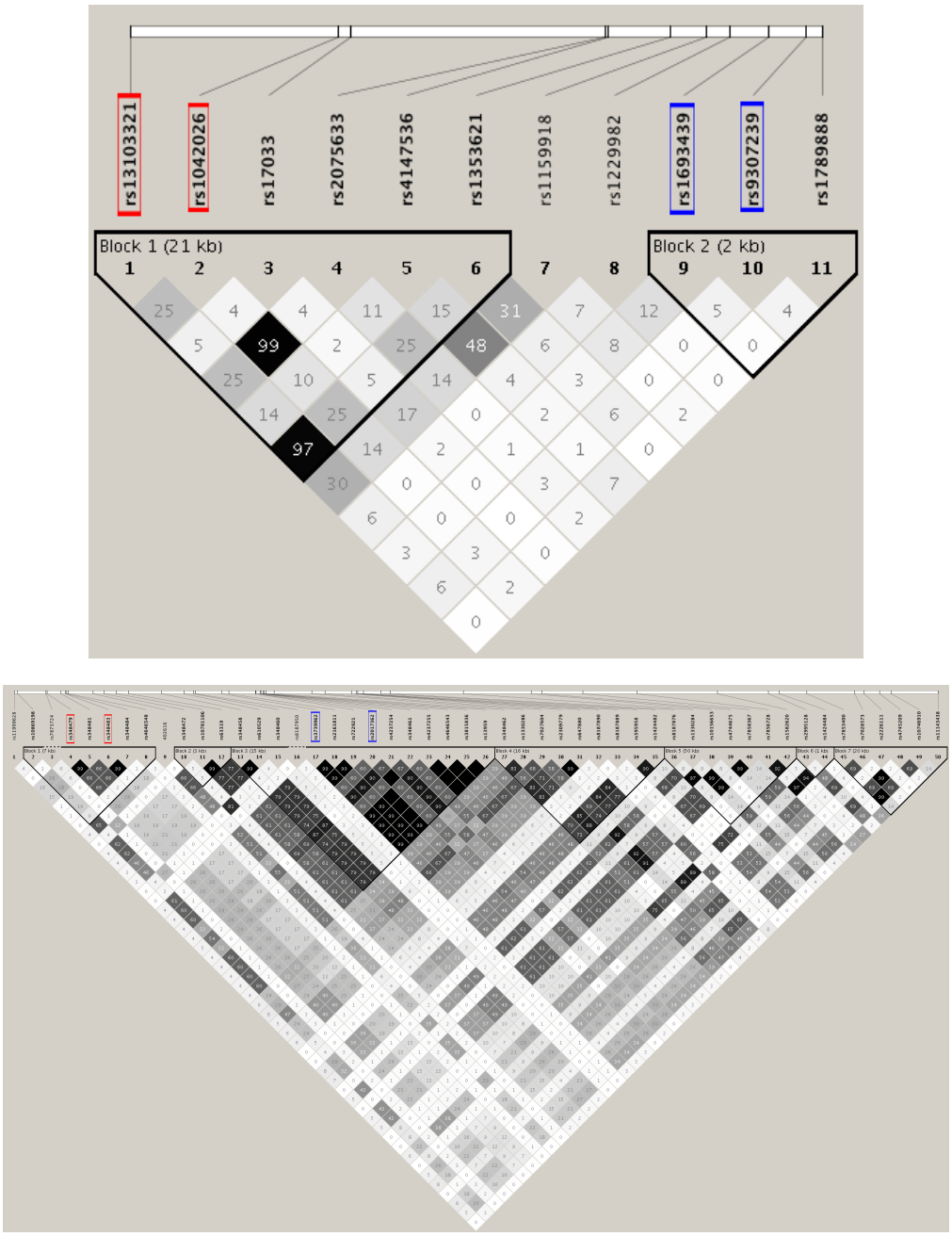

Figure 3.1: The LD structure of the 11 SNPs of ADH1B and the 50 SNPs of ALDH1A1 in the MCTFR parental sample. The darker a square is, the higher R^2 between the two SNPs. The causal SNPs used in the simulations are markered in red (with interaction β_1) or blue (with interaction β_2)

0.14, respectively. Note that our combination test iSeq-aSum-min was protective against the inflated type I error of iSeq-aSum-G. The iSeq-aSum-min did not become inflated until the correlation was greater than 0.18. This simulation study demonstrated that even in the presence of G-E dependence, the methods that incorrectly model interaction give invalid p-values only in the presence of a substantially large genetic main effect. Furthermore, unless the causal SNPs are very correlated with the environment, our combination test maintains valid type I error even when there is a large genetic main effect.

Simulation 2 For the rest of our simulations, we assume that we have G-E independence ($\phi = 0$ in Equation 3.9). To evaluate the power of the methods to detect G-E interaction, we considered two basic scenarios for G-E interaction: (1) $\beta_1 = \beta_2$ or (2) $\beta_1 = -\beta_2$ in Equation 3.10. In each scenario, we varied β_1 from 0 to 0.2. We have also set the individual effect sizes of the causal SNPs equal to $\alpha_3 = 0.2$. The scenarios where we have (1) $\beta_1 = \beta_2$ or (2) $\beta_1 = -\beta_2$ can be seen in the left and right panels of Figure 3.3, respectively.

For scenario (1) where all of the non-null interactions are deleterious, iSeq-aSum-G performed best over the entire range of $\beta_1 = \beta_2$ for ADH1B. The relevant SNPs for interaction were captured by iSeq-aSum-G through the main effect only model and it gained power from using a chi-square test with one df. However, the burden test did not perform variable selection and thus did not perform very well here because of summing over all the SNPs including null variants. For the larger ALDH1A1 gene, we found that iSeq-aSum-I and the minP test had the worst power compared to the other methods, which all performed similarly. Here, the iSeq-aSum-I approach searched over a large number of possible allocations and hence lost power for ALDH1A1. Because ALDH1A1 has a moderate to high amount of LD, which produces more non-null effects, iSeq-aSum-G and the burden test can more effectively capture the interaction with a one df test. Meanwhile, GESAT and aSPU performed very similarly for both of the genes in scenario (1). However, these tests are best used in presence of many null variants.

For the scenario (2), we allowed half of the non-null interactions to be deleterious while the other half were protective. In this type of situation where there are effects with different directions, GESAT and aSPU typically perform well. However, we found

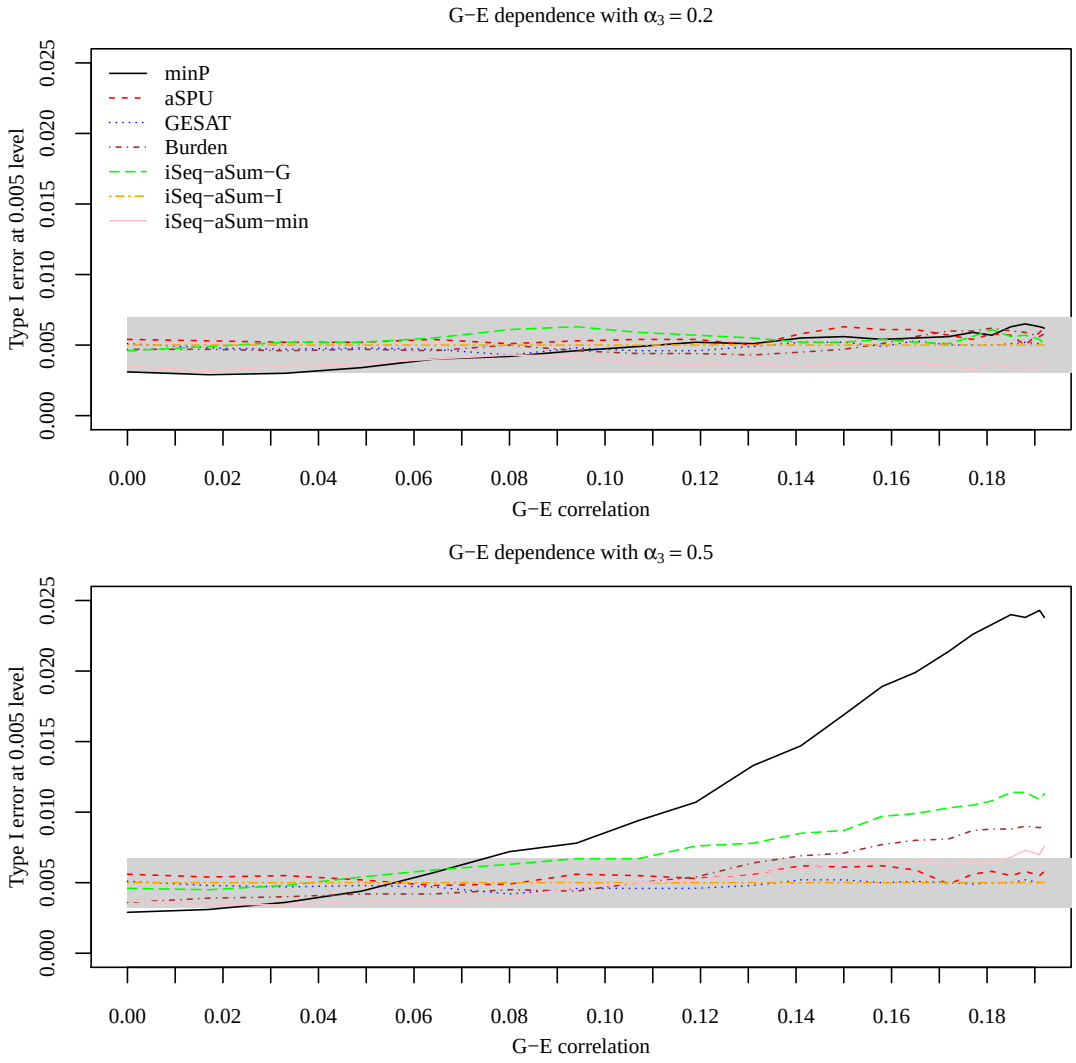

Figure 3.2: Simulation 1. Type I error of the methods for ADH1B as the amount of G-E dependence is increased. The G-E dependence is measured by the correlation between the burden summary measure and the environment. The grey bounds drawn at $0.005 \pm Z_{0.0025}\sqrt{\frac{(0.005)(0.995)}{10000}}$ denote Type I error estimates that are within sampling error.

that iSeq-aSum-I also captured such interaction effects and performed as well as GESAT and aSPU in ADH1B. Additionally, iSeq-aSum-I had much higher power than any other method for ALDH1A1 because the LD created a situation where there were not many null interactions that would give GESAT and aSPU an advantage over iSeq-aSum-I. As expected, the burden test and iSeq-aSum-G suffered extreme power loss for both genes due to not capturing the directionality of interaction from the main effect model. Finally, we saw that by combining iSeq-aSum-G and iSeq-aSum-I, iSeq-aSum-min performed well for both genes as shown in Figure 3.3.

Simulation 3 For our last simulation, we assessed the impact of the genetic main effect α_3 on the power to detect interaction for all of the methods. The iSeq-aSum-G test and therefore iSeq-aSum-min rely on the presence of a genetic main effect of the SNPs interacting with the environment. Simulation 2 used a moderately sized genetic main effect for the causal SNPs. With this main effect set to $\alpha_3 = 0.2$, the individual heritability estimates of the causal SNPs in both ADH1B and ALDH1A1 was 0.4% and 1.4% respectively. However, with smaller values of α_3 iSeq-aSum-G may fail to capture the interaction. To assess this, we fixed $\beta_1 = \beta_2 = 0.14$ for ADH1B and $\beta_1 = \beta_2 = 0.06$ for ALDH1A1 and varied α_3 from 0 to 0.4. In Figure 3.4, if there was no genetic main effect ($\alpha_3 = 0$), iSeq-aSum-G lost power. However, as we increased α_3, iSeq-aSum-G and thus iSeq-aSum-min quickly gained power to detect interaction. The iSeq-aSum-G test had good power to detect interaction for $\alpha_3 = 0.1$ and greater which corresponded to 0.2% and 0.4% of the total trait variation for ADH1B and ALDH1A1 respectively.

Another interesting finding from Figure 3.4 is the power loss of the burden and minP test as α_3 increases. While this happened for both genes, it was more apparent for ALDH1A1. To explain this, it is important to note that the burden and minP test incorrectly model the main effects. Incorrect specification of the main effect can lead to power loss for the interaction test. We show how the minP test can lose power due to incorrect specification of the main effect in Appendix A.3. However the power loss was not substantial within the realistic ranges of main effect for individual SNP. A similar argument can be made for the power loss of the burden test. On the other hand, iSeq-aSum-G assumes equal main effect sizes and the same direction of effects for the causal SNPs. In Figure 3.4, iSeq-aSum-G did not lose power for large values of α_3 in

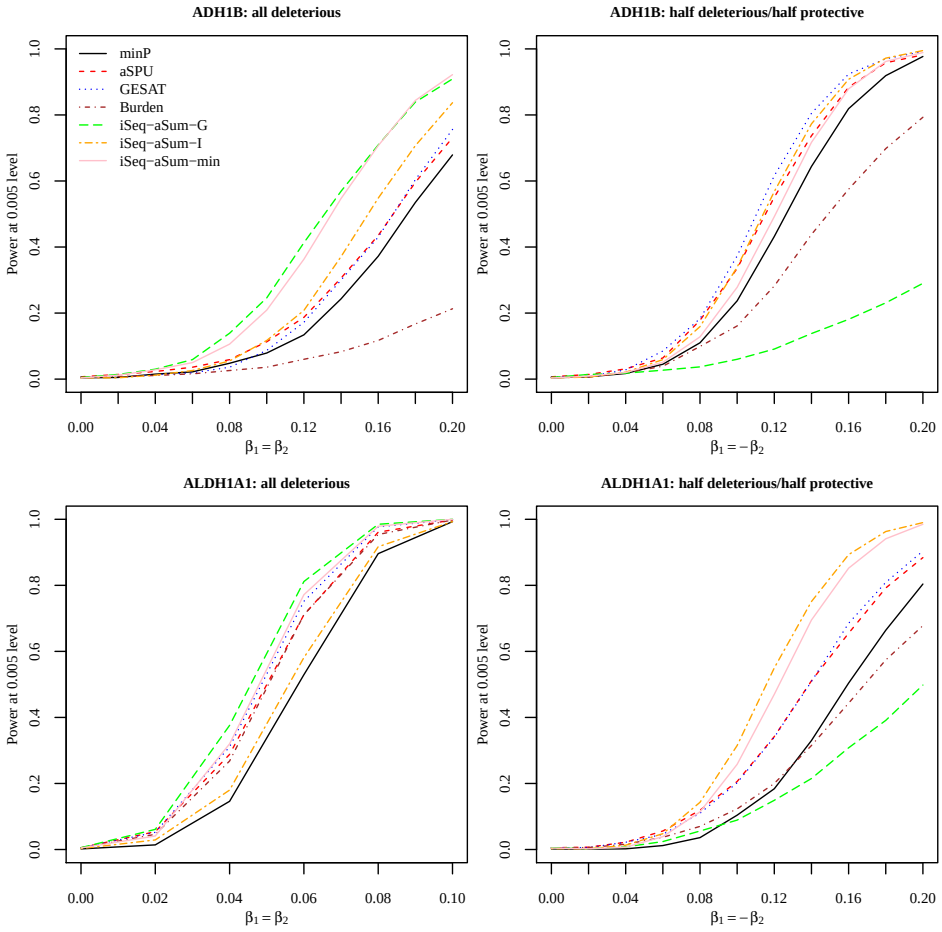

Figure 3.3: Simulation 2. Comparison of power under two scenarios for two different genes (Top panels: ADH1B, Bottom panels: ALDH1A1). Power curves for $n = 1000$ at $\alpha = 0.005$ level of significance with main effects present ($\alpha_3 = 0.2$) and G-E independence ($\phi = 0$). Right panels: half non-zero βs are equal while the other half are equal in the opposite direction ($\beta_1 = -\beta_2$).

Figure 3.4 as iSeq-aSum-G correctly modeled the main effects. The simulation model here assumed equal main effects and same directionality for the causal SNPs.

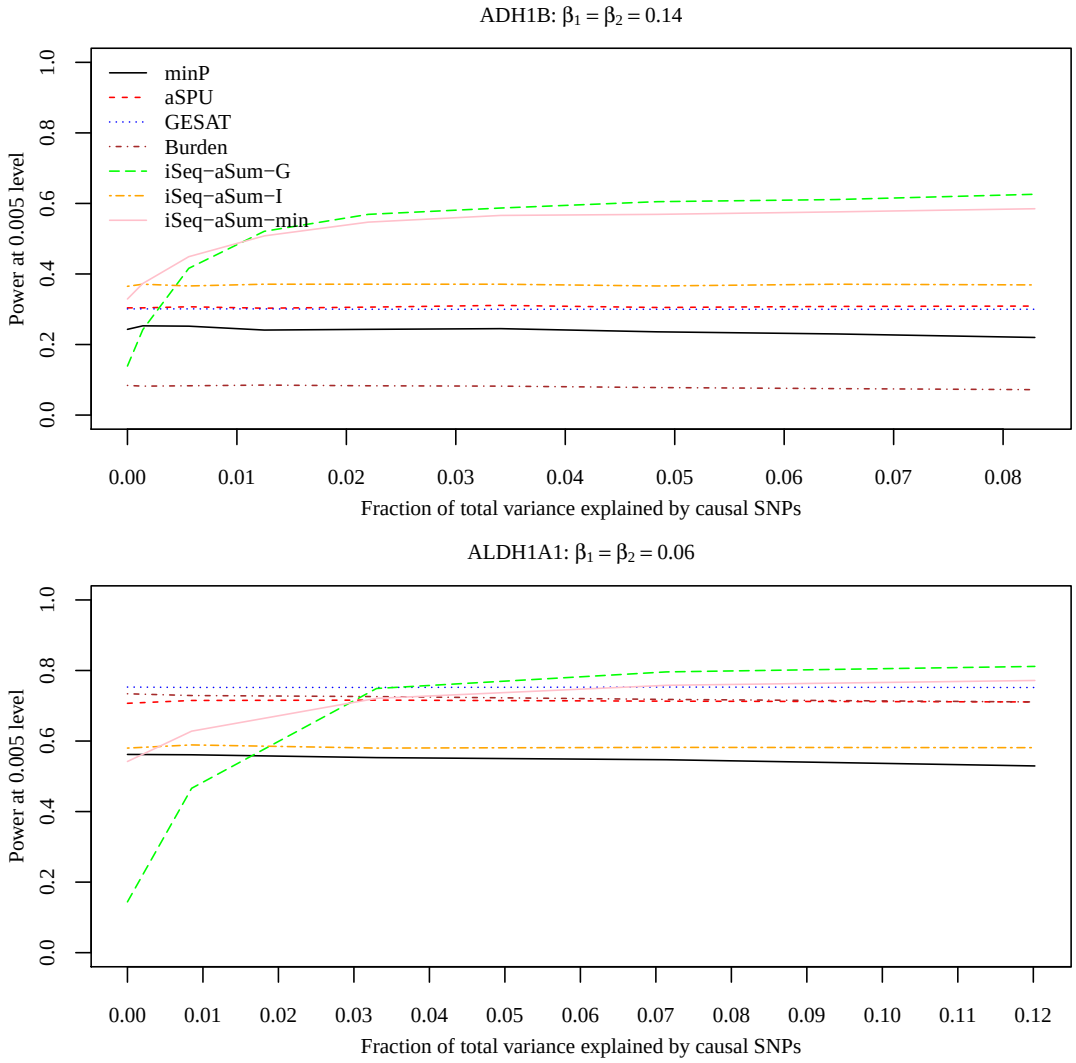

Figure 3.4: Simulation 3. Power of detecting G-E interaction with $\beta_1 = \beta_2 = 0.14$ for ADH1B and $\beta_1 = \beta_2 = 0.06$ for ALDH1A1 while varying main effect of the four causal SNPs (α_3) with G-E independence ($\phi = 0$). Power curves for $n = 1000$ at $\alpha = 0.005$ level of significance.

3.4 Minnesota Center for Twin and Family Research

The MCTFR aims to identify factors that contribute to the development of psychological outcomes such as substance use disorders (Miller et al., 2012). Here, we are interested in studying the factors that may contribute to alcohol consumption. To quantify alcohol consumption, we used composite quantitative clinical phenotypes derived from a hierarchical factor analytic approach as described in Hicks et al. (2011). Previous studies of the MCTFR data have estimated alcohol consumption to be highly heritable with about 50% to 80% of the phenotype explained by the genetic variation in the MCTFR nuclear families with parents and biological offspring (McGue et al., 2013). However, genetic association studies with common and rare variants have explained little of the estimated heritability in this study (McGue et al., 2013; Vrieze et al., 2014). Some of the missing heritability may be due to environmental factors moderating the genetic main effect. Hicks et al. (2009) found evidence of G-E interaction with family climate through biometric modeling of the MCTFR dataset. The goal of this analysis is to implement the different G-E interaction tests we considered in Section 3.2 to detect G-E interaction between 'family climate' and selected candidate genes for alcoholism in MCTFR study. Due to the limited sample size of MCTFR, we only selected the ten most studied genes of relevance to alcohol abuse from Olfson and Bierut (2012) in order to gain power to detect interactions. The genes studied here are: GABRA2, ADH1B, ADH1C, SLC6A3, SLC6A4, OPRM1, CYP2E1, DRD2, ALDH2, and COMT. We use $\alpha = 0.05/10 = 0.005$ to identify significant findings.

To obtain the SNP-set for each candidate gene, SNPs genotyped using the Illumina Human660W-Quad Array chip were mapped to the closest gene using an NCBI Build 36 annotation file. Our environmental factor of interest, 'family climate' score, was computed separately for mothers and fathers using the scales of the Parent Environment Questionnaire (PEQ) (Elkins et al., 1997). Specifically, the PEQ consists of 5 scales: Conflict with Parent, Involvement with Parent, Regard for Parent, Regard for Offspring, and Structure. Parents completed the PEQ twice, reporting on their relationship with the target offspring as well as the other parent's relationship. As shown by Elkins et al. (1997), the PEQ scales are moderately intercorrelated and when factor analyzed a single factor emerges. This factor score, averaged over the multiple raters, was used

as the 'family climate' score for each parent. The first four principal components were included as covariates in all of the models to adjust for population stratification. We also included age as a covariate in the model. For this analysis, we considered only the unrelated individuals of the MCTFR: the parents. We separately analyzed the fathers and mothers because they were correlated due to shared environment. The correlation for alcohol consumption phenotype between mothers and fathers was 0.44, and the main effect of family climate on alcohol consumption was significantly different between the sexes. This finding is consistent with previous findings that the disease etiology of alcoholism differs by sex (Prescott, 2002; Hardie et al., 2008). Our analyses used the 1,446 fathers and 1,756 mothers who passed quality control filtering and had complete data for all variables described above.

For the MCTFR analyses, we compared the performances of the minP test, aSPU, GESAT, iSeq-aSum-G, iSeq-aSum-I, and iSeq-aSum-min. We used Bootstrap Strategy 2 to compute p-values for both iSeq-aSum-I and iSeq-aSum-min. To validate the approximate p-values for iSeq-aSum-I, we used Bootstrap Strategy 1 with $B = 10000$ and found little difference between these strategies. The p-value for iSeq-aSum-min calculated using Bootstrap Strategy 1 was smaller than the approximate p-value calculated by Bootstrap Strategy 2 because the Bonferroni correction was conservative. The correlation between the burden summary measure and family climate ranged from -0.03 to 0.03. A linear regression of the burden summary measure on the family climate score showed no evidence of G-E dependence. Additionally, our parametric bootstrap p-values for iSeq-aSum-G and the burden test agreed with the p-value derived from the asymptotic chi-square distribution with one df.

In Table 3.1, we reported the genes identified by at least one method to have significant interaction with family climate. For simplicity, we only reported p-values for iSeq-aSum-I and iSeq-aSum-min as calculated by Bootstrap Strategy 2. Three out of the ten genes showed evidence of interaction with family climate for parental alcohol consumption. The GABRA2, ADH1C, and DRD2 genes encode for a neurotransmitter, an alcohol metabolite, and a dopamine receptor, respectively. Our interaction summary method iSeq-aSum-I identified all three significant genes, whereas GESAT and the burden test identified one each. Our combination approach, iSeq-aSum-min was significant for two genes. However, if we used the less conservative strategy (Bootstrap Strategy 1)

to calculate the p-values, iSeq-aSum-min was significant for all three genes. Note that while the burden test and iSeq-aSum-I are both significant for DRD2, iSeq-aSum-G has a very large p-value. By looking at Figure 3.4, our simulations studies indicate that the variants with interaction in DRD2 probably have weaker main effects, which were not captured by iSeq-aSum-G. For ADH1C, however, iSeq-aSum-G is almost significant while the burden test is not significant. The variable selection feature of iSeq-aSum-G possibly gave the test an advantage over the burden summary measure.

Table 3.1: Genes with significant interaction with family climate in the MCTFR analysis. The correlation between the environment and the burden summary measure found by iSeq-aSum-G is reported and showed no evidence of G-E dependence. All of the methods are compared for each gene with significant p-values marked in bold.

Gene	**ADH1C**	**GABRA2**	**DRD2**
# SNPs	8	26	44
Parent	Father	Mother	Mother
G-E correlation	0.030	0.006	-0.025
p-value	0.251	0.792	0.290
minP test	0.0070	0.0689	0.1138
aSPU	0.06294	0.2957	0.0160
GESAT	**0.0026**	0.1010	0.03616
Burden	0.5971	0.1866	**0.0018**
iSeq-aSum-G	0.0090	0.1245	0.6567
iSeq-aSum-I	**0.0025**	**0.0022**	**0.0032**
iSeq-aSum-min	0.0050	**0.0044**	0.0064

3.5 Discussion

In this paper, we have extended the sequential algorithm of Basu and Pan (2011) to test for G-E interactions through iSeq-aSum-G and iSeq-aSum-I. These two tests can further be combined using iSeq-aSum-min. The advantage of this combination test is to significantly improve the power for G-E interaction detection while largely controlling for the type I error even in presence of G-E dependence. We derived the asymptotic

distribution of iSeq-aSum-I for uncorrelated SNPs and used a parametric bootstrap to estimate the distribution in the presence of LD. We used this computationally efficient parametric bootstrap approach to estimate the p-value of iSeq-aSum-min approach as well. In our simulation studies, this combination test showed great power under various alternative simulation models and maintained correct type I error even under strong G-E dependence. In general, iSeq-aSum-G performed well if the interactions were captured through the main effects, while iSeq-aSum-I performed well if there were directional effects that could not be captured by a main-effect model. By combining these two tests, iSeq-aSum-min was able to take advantage of the strengths of both iSeq-aSum-G and iSeq-aSum-I.

It has been already established that the minP test and tests based on summary measures, including iSeq-aSum-G, are biased under G-E dependence (Lin et al., 2013, 2015). However, in our simulations we demonstrated that the impact of this G-E dependence only occurs in extreme circumstances. With small to moderate genetic main effects, our simulations showed little influence on the type I error in presence of G-E dependence. Additionally, the type I error for iSeq-aSum-min, which incorporates iSeq-aSum-G, was not affected unless there was a strong genetic main effect and strong G-E dependence.

There was not a single test that was uniformly most powerful under different alternative simulation models. It is important to note that the two genes used in our simulations had few interactions with no effect. If there are many null effects, it has been well established that the aSPU and GESAT tests will perform well (Basu and Pan, 2011; Lin et al., 2013, 2015). ADH1B had only eleven SNPs in low LD and four of these variants had interaction in our simulations. ALDH1A1 had high LD among the variants and hence created situations where there were not many null variants in our simulation study. Hence, we did not see much power advantage of aSPU or GESAT over other methods. The simulation setup did not favor the minP test as well. The minP test would perform well if there is a strong contribution from a single SNP. The minP test will lose power if multiple SNPs moderately contribute to the interaction and have sizable main effects (Appendix A.3).

We also tested these methods' power to detect interactions using real data from the MCTFR parent cohort. In this case, we found that iSeq-aSum-I detected three genes. Similarly, iSeq-aSum-min detected two of these three genes, while narrowly missing the

other one. However, by using Bootstrap Strategy 1, the full parametric bootstrap, this combination test performed as well as iSeq-aSum-I. The burden test and GESAT were able to identify only one of the three genes.

The proposed methods have some potential limitations. We suggest estimating the approximate chi-square distribution for iSeq-aSum-I using Bootstrap Strategy 2. While Bootstrap Strategy 2 is significantly more computationally efficient than Bootstrap Strategy 1, Bootstrap Strategy 2 requires more computation than the other methods. However, parallel computing can reduce the time to compute these p-values because each parametric bootstrap sample is independent. Another limitation is that aSPU and iSeq-aSum-I estimate all of the main effects from Model 3.1 and may encounter estimation issues for larger numbers of SNPs. Using a ridge penalty on the main effect similar to GESAT can mitigate this problem. Another strategy is to avoid fitting all of the main effects, as is done by the minP test and summary measure-based tests. In doing so, however, these tests lose power to detect interactions as the main effect increases. They can also have inflated type I error in the presence of G-E dependence.

Here, we have focused on gene-based G-E interaction tests for only unrelated subjects. However, the MCTFR study is focused on detecting potential G-E interactions on the entire twin sample. We are currently extending our proposed test to consider family data.

Chapter 4

Score tests for gene-environment interaction in family studies using linear mixed models

4.1 Introduction

Over the past two decades, researchers have concentrated on studying the genetic contribution to complex diseases using genome-wide association studies (GWAS). However, the single nucleotide polymorphisms (SNPs) identified in GWAS only explain a small proportion of the disease heritability. This may be because the genetic risk of the SNPs is modified by environmental factors. Thus, understanding the interplay between genes and environments can further help us understand complex diseases. Many examples of gene-environment (GxE) interaction have been found for a variety of diseases (Hunter, 2005). Identification of these interactions is important for understanding underlying disease etiology and developing disease prevention and intervention strategies. Here, we aim to identify GxE interactions in a family study by testing interactions between a group of SNPs from a candidate gene and a set of environmental factors.

One strategy to identify GxE interaction between a candidate gene and a set of correlated environmental factors is to test the interaction of each SNP and each environmental factor separately and subsequently apply a multiple testing correction. A

severe limitation of this approach is that the type I error rate can be inflated if the SNPs and environments are correlated (Lin et al., 2013). Another limitation is that single marker tests do not incorporate the possible joint effects among the SNPs, the environments, and the interactions. By not taking advantage of possible joint effects, single marker tests of interaction can lose power (Lin et al., 2013; Coombes et al., 2016). Recently, several gene-based tests for interaction between a group of SNPs in a candidate gene and one environmental measure have been developed (Lin et al., 2013, 2015; Wang et al., 2015; Coombes et al., 2016). Most of these tests can be implemented using the score vector of the GxE interactions and its covariance.

In their current forms, however, these methods are not suitable for use with correlated subjects such as families. Using the current methods for GxE interaction without accounting for within family similarities can result in inflated type I error (Chen et al., 2013). It is important to extend these methods for family data because family studies can be very useful for testing for GxE interaction. By matching on genotype in these studies, the proportion of genotype-concordant, exposure-discordant pairs may be much higher than in studies with unrelated subjects (Thomas, 2010b; Yang and Khoury, 1997). Additionally, with the increasing emphasis on finding GxE interactions, there are many family studies available, such as the Framingham Heart Study (Dawber et al., 1951), the National Heart, Lung, and Blood Institute Family Heart Study (Higgins et al., 1996), and the STANISLAS cohort (Visvikis-Siest and Siest, 2008) which would require tools to analyze GxE interactions within the family framework.

Our work is motivated by data originating from the Minnesota Center for Twin and Family Research (MCTFR) study which investigates psychological outcomes such as substance use disorders (SUDs) (Miller et al., 2012). Addiction to several different substances including caffeine, nicotine, alcohol, cannabis, sedatives, stimulants, cocaine, and hallucinogens all appear to be moderately to strongly heritable, but genetic association studies with common and rare variants have explained little of the estimated heritability in this study (McGue et al., 2013; Vrieze et al., 2014). Using biometric models, the MCTFR has shown how particular environmental factors relate to substance abuse risk and interact with genetic risk for the twin sample (Hicks et al., 2011; Samek et al., 2016). Our methodological work in this paper aims to facilitate the detection of specific genes that are involved in this interaction.

Finally, the methods for GxE interaction mentioned above have only focused on interactions of a set of SNPs from a candidate gene with a single environmental factor. However, in the MCTFR study, we focus on a group of four correlated environmental factors that may interact with genes to influence alcohol consumption. One strategy to model multiple environmental factors is to separately test for interaction between the set of SNPs and each environmental factor. However, similar to the arguments made for analyzing correlated SNPs within the same gene together (Peng et al., 2009), including all of the environments and their interactions with the SNP-set in one model can increase power for detection of GxE interaction.

In this article, we use the score vector for the GxE interactions and its covariance within a linear mixed model (LMM) framework to extend and propose new tests of GxE interaction for family data. We either test one environment at a time or all at once in this model. Our simulations show that using all environments in one model increases the power to detect interaction. Within both LMM approaches, we also implement a ridge penalty on the genetic main effect using a random effect. This reduces the number of parameters we need to estimate which produces more powerful tests of GxE interaction. Our simulations show that we gain power as a result of using the ridge penalty. Using the resulting score vector of GxE interaction and its covariance, we extend the score test, gene-environment set association test (GESAT) (Lin et al., 2013), adaptive Sum of Powered Score tests (aSPU) (Pan et al., 2014), and the interaction test using a Sequential adaptive Sum (iSeq-aSum) (Coombes et al., 2016) from independent subjects to families. Additionally, we propose a generalization of iSeq-aSum using the family of powered score tests as first proposed by Pan et al. (2014). In fact, the resulting family of tests, the Sequential algorithm for the sum of powered score (Seq-SPU) tests, is equivalent to a weighted version of the SPU tests when weights are chosen using a sequential algorithm. While even-powered Seq-SPU tests perform similarly to their SPU counterparts, extensive simulations show odd-powered Seq-SPU tests can be much more powerful than the SPU equivalent when there are a mix of positive and negative interaction effects. As a result, the adaptive version of Seq-SPU, Seq-aSPU, performs better than aSPU in these cases. Finally, we study the performance of the methods using the MCTFR dataset. We perform a gene-based GxE interaction analysis on the twin cohort of the MCTFR to study how genes from select candidate genes interact with

a set of environmental factors to affect alcohol consumption. One gene, only identified by Seq-aSPU, is found to significantly interact with the set of environmental factors to influence alcohol consumption in the MCTFR dataset.

4.2 Methods

To set up the GxE interaction model for family data, assume we have M independent families with m_i individuals where $i = 1, \cdots, M$. For the j^{th} individual from the i^{th} family, let Y_{ij}, $\mathbf{G}_{ij} = (G_{ij1}, \cdots, G_{ijq})^T$, $\mathbf{E}_{ij} = (E_{ij1}, \cdots, E_{ijp})^T$, $\boldsymbol{X}_{ij} = (X_{ij1}, \cdots, X_{ijL})^T$ be the phenotype, the q minor allele counts for the common genetic variants from a candidate gene, which are standardized by their mean and standard deviation, the p environmental factors, and L covariates, respectively. Define $\mathbf{S}_{ij} = \mathbf{G}_{ij} \otimes \mathbf{E}_{ij} = (G_{ij1}\mathbf{E}_{ij}^T, \cdots, G_{ijq}\mathbf{E}_{ij}^T)^T$ to be the pq pairwise GxE interactions for the ij^{th} individual where $\otimes$ is the Kronecker product.

An approach to modeling dependent observations is to model the dependency structure with random effects. While it is possible to use random effects for discrete outcomes, estimation is much more difficult because the likelihood functions cannot be derived in closed form (Pinheiro and Chao, 2006). Therefore, we restrict our scope to continuous outcomes and focus on obtaining the score vector for the GxE interaction as well as its covariance in the linear mixed model (LMM) setting. We consider testing for GxE interaction among the q SNPs and the p environments in either one model with all of the pq pairwise interaction terms included as discussed in Section 4.2.1 or in p separate models each with q pairwise interaction terms included as discussed in Section 4.2.2.

4.2.1 Joint modeling of GxE interaction

To test for all pq interactions in one model, we use the following LMM:

$$Y_{ij} = \alpha_0 + \boldsymbol{X}_{ij}^T\boldsymbol{\alpha}_1 + \mathbf{E}_{ij}^T\boldsymbol{\alpha}_2 + \mathbf{G}_{ij}^T\boldsymbol{\alpha}_3 + \mathbf{S}_{ij}^T\boldsymbol{\beta} + a_{ij} + c_{ij} + e_{ij} \tag{4.1}$$

where $\alpha_0, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\beta}$ are the regression coefficients for the intercept, covariates, environmental factors, genetic variants, and GxE interactions, respectively. For Model 4.1, we are interested in testing the null hypothesis that there is no GxE interaction ($H_0 : \boldsymbol{\beta} = 0$). We refer to this strategy that includes all environmental factors in the model

as the Joint approach because it considers possible joint effects between environments and their interactions with the SNP-set.

To account for shared genetic effects within families, let $\boldsymbol{a}_i = (a_{i1}, \cdots, a_{im_i}) \sim \text{MVN}(\mathbf{0}, \mathbf{A}_i)$ where $\mathbf{A}_i = \sigma_A^2 \mathbf{K}_i$ and $\mathbf{K}_i$ is equal to two times the kinship matrix for the i^{th} family, where $[\mathbf{K}_i]_{r,s}$ is the probability that a gene is identical-by-decent for the r^{th} and s^{th} member of the i^{th} family. For example, if we have a family consisting of either monozygotic (MZ) or dizygotic (DZ) twins, the off-diagonal elements of $\mathbf{K}_i$ are 1 or 1/2, respectively. To account for shared environmental effects within families, let $\boldsymbol{c}_i = (c_{i1}, \cdots, c_{im_i})$ be a random intercept for family defined as $\boldsymbol{c}_i \sim \text{MVN}(\mathbf{0}, \mathbf{C}_i)$ where $\mathbf{C}_i = \sigma_C^2 \mathbf{J}_{m_i}$ and $\mathbf{J}_{m_i}$ is an $m_i \times m_i$ matrix of ones. While Model 4.1 includes some shared environments as fixed effects, there may still be unmeasured shared environments which can be accounted for by σ_C^2. Finally, we let $\boldsymbol{e}_i = (e_{i1}, \cdots, e_{im_i}) \sim \text{MVN}(\mathbf{0}, \mathbf{E})$ where $\mathbf{E} = \sigma_E^2 \mathbf{I}_{m_i}$. This term accounts for all other unshared effects. The covariance between different families is zero because different families are assumed to be independent. We refer to Model 4.1 as the ACE model because it splits the covariance within families into three parts: $\mathbf{A}$ = shared genetic effects, $\mathbf{C}$ = shared environmental effects, and $\mathbf{E}$ = unshared environmental effects (Falconer and Mackay, 1981). Note that depending on our family structures, σ_A^2 and σ_C^2 may not be identifiable. However, for our simulations and real data application to the MCTFR which contains MZ and DZ twins, these parameters are identifiable. In our simulations, we investigate the effect of not estimating σ_C^2 for twins. We refer to this as the AE model. In our simulations, we explore the consequences of failing to adjust for unmeasured shared environments with the AE model when our goal is test for GxE interaction.

A concern common in analyzing high dimensional genetic data is that Model 4.1 requires estimation of $K + p + q + 1$ main effects where q can be very large for some genes. This may cause estimation issues and impact our GxE interaction test. Lin et al. (2013) and Lin et al. (2015) have previously penalized genetic main effects using a ridge penalty to alleviate this issue. However, using a ridge penalty within an LMM framework can be computationally intensive. Instead, we re-cast the ridge penalization of the SNPs using a random effect by allowing $\boldsymbol{\alpha}_3 \sim \text{MVN}(\mathbf{0}, \sigma_G^2 \mathbf{I}_q)$ in Equation 4.1 (Hodges, 2013; Shen et al., 2013). Under this setup, the traditional ridge penalization parameter λ is equivalent to σ_E^2/σ_G^2. Further detail can be found in Appendix B.1. Now, we only need

to estimate one additional variance term. This formulation of $\boldsymbol{\alpha}_3$ has been previously been used in the prediction literature and $q \times \sigma^2_G$ can be interpreted as the proportion of variation explained by the SNPs (Speed and Balding, 2014; Yang et al., 2011). Here, it is used as a convenient tool to estimate the genetic main effects even when our SNP-set is very large.

To test the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, we use the score vector of $\boldsymbol{\beta}$ which can easily be derived as $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y} - \mathbf{1}\hat{\alpha}_0 - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_1 - \mathbf{E}\hat{\boldsymbol{\alpha}}_2)$ where $\mathbf{S} = (\mathbf{S}_{11} \cdots \mathbf{S}_{ij} \cdots \mathbf{S}_{Mm_M})^T$, $\boldsymbol{Y} = (Y_{11}, \cdots, Y_{Mm_M})^T$, $\boldsymbol{X} = (\boldsymbol{X}_{11} \cdots \boldsymbol{X}_{Mm_M})^T$, and $\mathbf{E} = (\mathbf{E}_{11} \cdots \mathbf{E}_{Mm_M})^T$. The estimated covariance is $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2_G \mathbf{G}\mathbf{G}^T + \hat{\sigma}^2_A \mathbf{K} + \hat{\sigma}^2_C \mathbf{C} + \hat{\sigma}^2_E \mathbf{I}$ where $\mathbf{G} = (\mathbf{G}_{11} \cdots \mathbf{G}_{Mm_M})^T$ and $\mathbf{K}$ and $\mathbf{C}$ are block diagonal with $\mathbf{K}_i$ and $\mathbf{J}_i$ on the diagonals, respectively. We use the `R` package `regress` to obtain our estimates of $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2)$ and $\hat{\boldsymbol{\Sigma}}$ under the $H_0 : \boldsymbol{\beta} = 0$. The Fisher Information of $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ is $\mathbf{I}(\boldsymbol{\theta}) = (\tilde{\boldsymbol{X}}|\mathbf{S})^T \hat{\boldsymbol{\Sigma}}^{-1} (\tilde{\boldsymbol{X}}|\mathbf{S})$ where $\tilde{\boldsymbol{X}} = [\mathbf{1}|\boldsymbol{X}|\mathbf{E}]$. $\mathbf{I}(\boldsymbol{\theta})$ can be partitioned into matrices $\mathbf{I}_{XX}$, $\mathbf{I}_{XS}$, $\mathbf{I}_{SX}$, and $\mathbf{I}_{SS}$ according to the dimensions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Thus, the covariance of the score vector of $\boldsymbol{\beta}$ is $\mathbf{V} = \mathbf{I}_{SS} - \mathbf{I}_{SX}\mathbf{I}^{-1}_{XX}\mathbf{I}_{XS}$.

We use the LMM score vector $\mathbf{U} = \mathbf{U}(\boldsymbol{\beta})$ with its covariance matrix $\mathbf{V}$ to extend current score tests of GxE interaction to families: the score test, aSPU (Pan et al., 2014), GESAT (Lin et al., 2013), and iSeq-aSum (Coombes et al., 2016). In similar fashion to Pan et al. (2014), we develop a larger family of tests called Seq-aSPU which can incorporate iSeq-aSum.

Score Test

Given $\mathbf{U}$ and $\mathbf{V}$, the score test can be calculated as $T_{sco} = \mathbf{U}^T\mathbf{V}^{-1}\mathbf{U}$ which has an asymptotic chi-square distribution with pq degrees-of-freedom (df) under H_0. However, if the number of variants q in a candidate gene is large, the score test can lose power to detect interaction due to its large df.

Adaptive Sum of Powered Score Tests

If we instead calculate our test statistic without using $\mathbf{V}$, Pan (2009) showed that $T_{ssu} = \mathbf{U}^T\mathbf{U}$ may more efficiently test for the combined effect of $\boldsymbol{\beta}$ because it uses fewer df. As an extension, Pan et al. (2014) proposed aSPU for testing for genetic main effects. This test was recently extended to tests of GxE interaction for independent

subjects (Coombes et al., 2016). Using the score vector $\mathbf{U}$, we can construct the GxE interaction SPU test for families as $T_{SPU(\gamma)} = \mathbf{1}^T\mathbf{U}^\gamma$ where $\mathbf{U}^\gamma = (U_1^\gamma, \cdots, U_{pq}^\gamma)$ for a set of integers $\gamma \geq 1$. As γ increases, the larger components of $\mathbf{U}$ are weighted higher. The null distribution of these SPU test statistics may be difficult to derive. However, under H_0, the score vector $\mathbf{U} \sim \mathcal{N}(0, \mathbf{V})$. Therefore, we can generate B copies of the null score vector by sampling from $\mathcal{N}(0, \mathbf{V})$ for which we calculate B copies of the SPU test $T_{SPU(\gamma)}^{(b)}$ where $b = 1, \cdots, B$. The p-value for a given γ is thus $P_{SPU(\gamma)} = (\sum_{b=1}^{B} I(|T_{SPU(\gamma)}^{(b)}| > |T_{SPU(\gamma)}|) + 1)/(B+1)$. Using the p-values for a set of γs, Pan et al. (2014) also proposes the aSPU test

$$T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)}.$$

where $\Gamma = \{1, 2, \cdots, 8, \infty\}$. With $\gamma = \infty$, this is similar to the UminP test by using the maximum value from the score vector. Using the same B copies of the null score vector, we can calculate the aSPU test statistic for each null score vector $T_{aSPU}^{(b)}$ and find the proportion of null aSPU test statistics that are smaller than our observed aSPU test statistic. Thus, the p-value for the aSPU test is $P_{aSPU} = (\sum_{b=1}^{B} I(T_{aSPU}^{(b)} > T_{aSPU}) + 1)/(B+1)$.

Here, the SSU test (SPU(2)) statistic is equivalent to extending the gene-environment set association test (GESAT) statistic as proposed by Lin et al. (2013) to family data. With $\gamma = 2$, the SPU test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is equivalent to the score test of $H_0 : \tau = 0$ when we define $\boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, \tau\mathbf{I}_{pq})$. However, rather than using a sampling method to calculate a p-value, we can use the characteristic function inversion method to calculate an asymptotic p-value (Lin et al., 2013). Under the null hypothesis, the variance of the residuals is

$$\text{var}(\boldsymbol{Y} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\alpha}}) = \hat{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\hat{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T = \mathbf{P}_0.$$

Thus, $\mathbf{U}^T\mathbf{U} \sim \sum_{k=1}^{q} \lambda_k \chi_{1,k}^2$ where λ_k are the eigenvalues of the matrix $\mathbf{S}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_0\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}$. The p-value is then computed analytically using the Davies method (Davies, 1980).

Sequential algorithm for aSPU

Another subset of the SPU test, SPU(1), is very similar to the Sum test. The Sum test uses a pooled regression estimate to model GxE interaction:

$$Y_{ij} = \alpha_0 + \boldsymbol{X}_{ij}^T\boldsymbol{\alpha}_1 + \mathbf{E}_{ij}^T\boldsymbol{\alpha}_2 + \mathbf{G}_{ij}^T\boldsymbol{\alpha}_3 + \beta_c \sum_{k=1}^{pq} d_k S_{ijk} + a_{ij} + c_{ij} + e_{ij} \tag{4.2}$$

where $d_k = 1$ for all k and β_c is the pooled GxE interaction regression estimate. The null hypothesis we wish to test is $H_0 : \beta_c = 0$. A score vector of β_c can be derived as

$$\begin{aligned} U(\beta_c) &= \left(\sum_{k=1}^{pq} d_k \mathbf{S}_k\right) \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y} - \mathbf{1}\hat{\alpha}_0 - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_1 - \mathbf{E}\hat{\boldsymbol{\alpha}}_2) \\ &= \sum_{k=1}^{pq} d_k \left(\mathbf{S}_k \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y} - \mathbf{1}\hat{\alpha}_0 - \boldsymbol{X}\hat{\boldsymbol{\alpha}}_1 - \mathbf{E}\hat{\boldsymbol{\alpha}}_2)\right) \\ &= \sum_{k=1}^{pq} d_k U_k = \boldsymbol{d}^T\mathbf{U} \end{aligned}$$

where $\boldsymbol{d} = (d_1, \cdots, d_{pq})^T = \mathbf{1}$ and $\mathbf{S}_k$ is the $\sum_{i=1}^{M} m_i \times 1$ vector for the k^{th} GxE interaction. Notice that the score vector for β_c is a linear combination of $\mathbf{U}$. The score vector $\mathbf{U} \sim \mathcal{N}(0, \mathbf{V})$ under the null hypothesis, thus, $\boldsymbol{d}^T\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{d}^T\mathbf{V}\boldsymbol{d})$. Therefore, the score test statistic of β_c is

$$T(\boldsymbol{d}) = (\boldsymbol{d}^T\mathbf{U})^2/(\boldsymbol{d}^T\mathbf{V}\boldsymbol{d}) \tag{4.3}$$

which has an asymptotic χ^2 distribution with one df under the null hypothesis.

The benefit of the Sum test is that it tests a single parameter β_c. Thus, it has low df and possibly increased power to detect interactions. However, a common issue with the Sum test is that it loses power if there are a combination of interactions with positive and negative effects. Instead of simply summing up the score vector using $\boldsymbol{d} = \mathbf{1}$, Coombes et al. (2016) adaptively sums the GxE interactions with a chosen allocation vector $\boldsymbol{d} = (w_1 s_1, \cdots, w_{pq} s_{pq})$ where w_k is a chosen weight and s_k indicates the directionality of the effect for the k^{th} interaction. We generally set $w_k = 1$ and set $s_k = 1$ or -1 which indicates an interaction effect is positive or negative. This test is referred to as iSeq-aSum. Coombes et al. (2016) and Basu and Pan (2011) showed that using an adaptively pooled effect estimate can avoid the power loss associated with

the Sum test when both positive and negative effects are present. However, if many interactions with no effect are pooled together with causal interactions, the regression estimate β_c will be pulled toward zero, which can result in a loss of power for both the Sum test and iSeq-aSum.

The SPU family of tests (Pan et al., 2014) can avoid losing power when there are many null interactions by increasing γ, so we use this strategy to propose a generalization of iSeq-aSum. To do this, we replace $\mathbf{U}$ in Equation 4.3 with $\mathbf{U}^\gamma$ to obtain the Seq-SPU test:

$$T(\boldsymbol{d}_\gamma, \gamma) = (\boldsymbol{d}_\gamma^T \mathbf{U}^\gamma)^2 / (\boldsymbol{d}_\gamma^T \mathbf{V} \boldsymbol{d}_\gamma) \tag{4.4}$$

where $\boldsymbol{d}_\gamma$ is allowed to vary for different γs. By taking the square root of $T(\boldsymbol{d}_\gamma, \gamma)$, it is easy to see that this test is equivalent to the weighted version of the SPU test (Kim et al., 2014) with weights equal to $\boldsymbol{d}_\gamma(\boldsymbol{d}_\gamma^T \mathbf{V} \boldsymbol{d}_\gamma)^{-1/2}$. Notice that iSeq-aSum is equivalent to Seq-SPU(1). Also, it is clear that as γ increases, the denominator weight $\boldsymbol{d}_\gamma^T \mathbf{V} \boldsymbol{d}_\gamma$ will have less impact on the allocation. If there are many interactions with no effect, which causes Seq-SPU(1) to lose power, in concordance with the SPU tests, we increase γ to avoid power loss. To find the optimal $\boldsymbol{d}_\gamma$ for a given γ, we proceed through the following sequential algorithm:

1. Initialize $\boldsymbol{d}_\gamma = \mathbf{1}$
2. `for` k in $1 : pq$
    - Set $d_{\gamma,k} = -1$ or 1 corresponding to the allocation that maximizes $T(\boldsymbol{d}_\gamma, \gamma)$

To compute p-values, we first generate B copies of the null score vector as before and find the optimal allocation $\boldsymbol{d}_\gamma^{(b)}$ for a given γ. We then calculate $T(\boldsymbol{d}_\gamma^{(b)}, \gamma)^{(b)}$ for each $\gamma \in \Gamma$ and $b = 1, \cdots, B$. The Seq-SPU(γ) p-values are computed as $P(\gamma) = (\sum_{b=1}^{B} I(|T(\boldsymbol{d}_\gamma^{(b)}, \gamma)^{(b)}| > |T(\boldsymbol{d}_\gamma, \gamma)| + 1)/(B+1)$. The Seq-aSPU test statistic is calculated as $\min_{\gamma \in \Gamma} P(\gamma)$. The p-value for Seq-aSPU can be calculated as before for aSPU.

While we perform a search over the entire set of $\Gamma = \{1, \cdots, 8, \infty\}$, we expect that only odd-valued γs will show power gain in comparison to their SPU counterparts because these γs are susceptible to power loss if there are a mix of positive and negative effects.

4.2.2 Univariate modeling of GxE interaction

An alternative to jointly testing for GxE interaction between the SNPs and all of the environments at once is to test for GxE interaction using each environment separately. For the k^{th} environment, we use the following LMM:

$$Y_{ij} = \alpha_0 + \boldsymbol{X}_{ij}^T\boldsymbol{\alpha}_1 + E_{ijk}\alpha_{2,k} + \mathbf{G}_{ij}^T\boldsymbol{\alpha}_3 + \mathbf{S}_{ijk}^T\boldsymbol{\beta}_k + a_{ij} + c_{ij} + e_{ij} \tag{4.5}$$

where $\mathbf{S}_{ijk} = \mathbf{G}_{ij} \times E_{ijk}$ for the j^{th} individual for the i^{th} family. For Model 4.5, we are interested in testing the null hypothesis $H_0 : \boldsymbol{\beta}_k = 0$ for all $k = 1, \cdots, p$. To obtain the score vector and covariance for each $\boldsymbol{\beta}_k$, we follow the same steps as in Section 4.2.1. For each environment, we use the LMM score vector of its interaction with the SNPs and the covariance matrix to calculate p-values for the Score, SPU, and Seq-SPU tests as outlined in Section 4.2.1. We obtain the p-value for the test of $H_0 : \boldsymbol{\beta}_k = 0$ for all $k = 1, \cdots, p$ by applying a Bonferroni correction to the minimum p-value of the p tests. We refer to this testing procedure as the MinP approach.

4.3 Results

We first compared through simulation studies the performance of the different methods to test for GxE interaction between a SNP-set from a candidate gene and a set of environments for a twin dataset. We generated datasets with 400 MZ and 250 DZ twin pairs using the genotype and environmental data from the MCTFR study described in Section 4.4. This sample size is chosen to mimic the sample size of our stratified analysis in Section 4.4. To preserve the correlation structures for the SNPs and environments in the MCTFR, we jointly sampled the SNPs, environments, and sex from the twins with complete data in the MCTFR dataset. We selected a candidate gene for alcoholism (Olfson and Bierut, 2012), ADH1B, which has 11 SNPs in low LD genotyped for the Illumina Human660W-Quad Array chip. Approximately 46% of the twins are male. Figure 3.1 shows the LD structure for these genotyped SNPs. The four environment scores sampled for our simulations are approximately normal with mean zero and standard deviations ranging from 0.77 to 0.89. Higher scores for the environments indicate greater risk for developing alcoholism. The correlation between pairs of environments ranges from 0.25 to 0.41. The environments are also correlated within twin pairs with

correlations ranging from 0.63 to 1. The genes and environments were correlated less than 0.03 which indicates that there is not gene-environment correlation.

We varied the number of causal SNPs Q in ADH1B and the proportion of total variance explained by these SNPs (R^2_G). We assumed that these SNPs only interacted with one environment and varied the proportion of total variance explained by the interactions (R^2_S). There are no other GxE interactions in our model. With these selected SNPs and interactions specified, we let

$$\text{var}(Y) = 1 = R^2_{sex} + R^2_E + R^2_G + R^2_S + \sigma^2_A + \sigma^2_C + \sigma^2_E \tag{4.6}$$

where $R^2_{sex} = 0.008$, $R^2_E = 0.35$, $\sigma^2_A = 0.3$, $\sigma^2_C = 0$, and $\sigma^2_E = 1 - R^2_{sex} + R^2_E + R^2_G + R^2_{GE} + \sigma^2_A + \sigma^2_C$ are the proportion of total variance explained by sex, four environments, genetic similarity, unmeasured shared environment, and error, respectively. For the i^{th} set of twins, we used the variance explained by each predictor, the male indicator $\boldsymbol{X}_i = (X_{i1}, X_{i2})$, four environments $\mathbf{E}_{i,k} = (E_{i1,k}, E_{i2,k})$ for $k = 1, \cdots, 4$, and the minor allele counts for the causal SNPs $\mathbf{G}_{i,k} = (G_{i1,k}, G_{i2,k})$ for $k = 1, \cdots, Q$ to simulate the phenotype $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$ as

$$\boldsymbol{Y}_i = b_0\mathbf{1} + b_{male}\boldsymbol{X}_i + b_E\sum_{k=1}^{4}\mathbf{E}_{i,k} + b_G\sum_{k=1}^{Q}\mathbf{G}_{i,k} + b_S\sum_{k=1}^{Q}d_k\mathbf{G}_{i,k}\cdot\mathbf{E}_{i,1} + \boldsymbol{a}_i + \boldsymbol{c}_i + \boldsymbol{e}_i \tag{4.7}$$

where $\cdot$ is the dot product, $b_0 = 0$, $b_{sex} = \sqrt{R^2_{sex}/(0.46(0.54))}$, $\boldsymbol{a}_i \sim \mathcal{N}(0, \sigma^2_A\mathbf{K}_i)$ where the off-diagonal elements of $\mathbf{K}_i$ are either 1 or 1/2 for MZ or DZ pair, respectively, and $\boldsymbol{e}_i \sim \mathcal{N}(0, \sigma^2_E\mathbf{I}_2)$. We assumed that each causal SNP had the same effect b_G. We set

$$b_G = \sqrt{\frac{R^2_G}{\text{var}\left(\sum_{k=1}^{Q}G_k\right)}} = \sqrt{\frac{R^2_G}{\sum_{k=1}^{Q}\text{var}(G_k) + \sum_{k\neq l}\text{cov}(G_k, G_l)}} \tag{4.8}$$

where G_k for $k = 1, \cdots, Q$ are correlated binomial random variables representing each causal SNP. We estimated the variance and covariance terms in Equation 4.8 using only one twin from each pair from the MCTFR data. With this setup, the causal SNPs collectively explain R^2_G of the total variance. We also determined the values of b_E and b_S in this way. For the causal interactions, we set $\boldsymbol{d} = (d_1, \cdots, d_Q)$ where $d_k = 1$ or -1 so that the interactions either have all the same directional effect or half of the interactions are positive and the other half are negative.

We are interested in testing the null hypothesis that there are no GxE interactions. To estimate the type I error for testing GxE interaction, we generated 10,000 null datasets with $R_S^2 = 0$ which corresponds to setting $b_S = 0$. To estimate the power for each method, we generated 1,000 datasets with $R_S^2 = 0.02$. The power and type I error were calculated at an $\alpha = 0.05$ or 0.01 level. To compute p-values for the SPU and Seq-SPU tests, we used $B = 1000$ to sample the null score vector. For each simulated dataset, we either tested for GxE interaction using the MinP approach as described in Section 4.2.2 or the Joint approach as described in Section 4.2.1.

In Simulation 1, we assessed the performances of the ACE and AE models and compared the power to detect interaction for the MinP and Joint approach. In Simulation 2, we assessed the impact of including a ridge penalty for the ACE model. Finally, we compared the performances of the SPU and Seq-SPU tests using the Score test as a control in Simulation 3.

4.3.1 Simulation 1

To compare the performances of the ACE and AE models for the MinP and Joint approaches, we set $Q = 2$, the fraction of total variance explained by the SNPs as $R_G^2 = 0.005$, and one interaction to be positive and the other negative. For Model 4.7, $\sigma_C^2 = 0$ for unmeasured shared environments. However, the MinP approach models each environment separately. Consequently, the environments not included in the test of interaction are "unmeasured" and resulted in $\hat{\sigma}_C^2 = 0.13$ for the ACE model in Table 4.1. Meanwhile, the AE model increased $\hat{\sigma}_A^2$ by the same amount to account for the environments that were left out. The other variance components were similar among the ACE and AE models and both of these models maintained type I error. Because we have used only twin data, specifying the A component as approximately A+C results in neglible difference between the ACE covariance structure and the AE covariance structure (See Appendix B.2 for more details). When we used the Joint approach, there was no difference between the ACE and AE models because the true σ_C^2 was set to zero in Equation 4.7.

We next compared power between the MinP and Joint approaches in Table 4.1. While only one environment has interaction with the causal SNPs, we have correlated environments and thus correlated GxE interactions which may make the Joint approach

more powerful than the MinP approach. Due to the large increase in df for the Score test (11 df $\rightarrow$ 44 df), the Score test's power for the MinP approach was higher than the Score test's power for the Joint approach with $\alpha = 0.01$. Alternatively, by using tests with smaller dfs, the power for aSPU and Seq-aSPU for the Joint approach was much higher than the MinP approach's power for either choice of α-level. For simplicity, the rest of our simulations use the ACE model with the MinP approach. Alternative model choices did not affect our conclusions for Simulations 2 and 3.

Table 4.1: **Simulation 1.** Type I error and power comparison of AE and ACE models for using environments separately or all together. Analyses were simulated using the ADH1B gene with two causal SNPs and interactions in opposite directions. The interactions explain 2% of the total variance of the simulated phenotype. $\hat{\sigma}^2_G$ is multiplied by 11 so that it can be interpreted as the proportion of variance explained by the 11 SNPs in the simulation.

		MinP Approach		**Joint Approach**	
	Model:	**ACE**	**AE**	**ACE**	**AE**
Mean (SD)	$11 \times \hat{\sigma}^2_G =$	0.005 (0.006)	0.005 (0.006)	0.005 (0.006)	0.005 (0.006)
of Variance	$\hat{\sigma}^2_A =$	0.313 (0.095)	0.444 (0.03)	0.253 (0.061)	0.293 (0.027)
Parameter Est.	$\hat{\sigma}^2_C =$	0.126 (0.083)	-	0.037 (0.049)	-
	$\hat{\sigma}^2_E =$	0.364 (0.028)	0.356 (0.026)	0.338 (0.025)	0.334 (0.024)
Type I error		**ACE**	**AE**	**ACE**	**AE**
$\boldsymbol{\alpha = 0.05}$	**Score**	0.0422	0.0402	0.0406	0.0420
	aSPU	0.0473	0.0452	0.0496	0.0524
	Seq-aSPU	0.0459	0.0451	0.0516	0.0519
$\boldsymbol{\alpha = 0.01}$	**Score**	0.0072	0.0071	0.0081	0.0074
	aSPU	0.0089	0.0081	0.0133	0.0120
	Seq-aSPU	0.0096	0.0081	0.0092	0.0122
Power		**ACE**	**AE**	**ACE**	**AE**
$\boldsymbol{\alpha = 0.05}$	**Score**	0.6169	0.6060	0.6190	0.6110
	aSPU	0.7362	0.7290	0.8960	0.8940
	Seq-aSPU	0.7653	0.7550	0.9080	0.9040
$\boldsymbol{\alpha = 0.01}$	**Score**	0.4333	0.4200	0.3690	0.3650
	aSPU	0.5918	0.5830	0.7750	0.7780
	Seq-aSPU	0.6078	0.5990	0.7850	0.7800

4.3.2 Simulation 2

To assess the potential gains of using a ridge penalty for the ACE model, we either fit the genetic main effects as fixed effects or as a random effect. We used $Q = 2$ with interactions in opposite directions. In Table 4.2, the number of SNPs in ADH1B (11) times $\hat{\sigma}^2_G$ correctly estimated the specified proportion of variance explained by the causal SNPs R^2_G. The A, C, and E variance components in Table 4.2 were very similar for all models, and there was no evidence of inflated type I error for any method or model. Due to the reduction in parameters to estimate, the model using ridge penalization was usually more powerful than the model that fitted the main effects as fixed effects regardless of the genetic main effect size; although the power difference is very small.

4.3.3 Simulation 3

For our last simulation, we compared the performances of aSPU and Seq-aSPU by studying the power of each family of tests over the entire $\Gamma = \{1, \cdots, 8, \infty\}$ set. We set the fraction of total variance explained by the SNPs at $R^2_G = 0.005$. We used the MinP approach with the ridge-penalized ACE model. Using $\alpha = 0.05$, Figure 4.1 shows the comparison between the powers of the Score, SPU, and Seq-SPU tests for two different scenarios: interactions with only one environment in the same or opposite directions.

For either scenario in Figure 4.1, we let the interactions of either two or four of the possible 11 SNPs ($Q = 2$ or 4) with the first environment explain 2% of the total variation in the simulated phenotype. When only two interactions had effect, higher-valued γs for either SPU or Seq-SPU had higher power because the nine interactions with no effect are weighted much smaller in comparison. Therefore, as the number of non-null interactions was increased to four, the lower-valued γs for both SPU and Seq-SPU increased in power because more interactions should be equally weighted. The biggest difference between aSPU and Seq-aSPU was seen in this case for either scenario because there are large differences between SPU and Seq-SPU with $\gamma = 1$ which corresponds to the Sum and iSeq-aSum tests, respectively.

In the first scenario in Figure 4.1a and 4.1b, we set the causal interactions to be in the same direction $\boldsymbol{d} = 1$. Because the effects of the causal interactions are in the same direction, using the sequential algorithm of iSeq-aSum to determine directional effects is

unneeded. Consequently, iSeq-aSum is less powerful than the Sum test and thus aSPU has higher power than Seq-aSPU. This difference is most evident when four out of 11 interactions had effect because the Sum test had much higher power than the higher df test, iSeq-aSum.

For the second scenario in Figure 4.1c and 4.1d, we set half of the interactions to be positive and the other half to be negative. As discussed in Section 4.2.1, this is a situation in which the even-valued γs in SPU are expected to perform much better than odd-valued γs. This pattern for the SPU tests held in our results regardless of the number of causal interactions and was especially true for $\gamma = 1$. Meanwhile, the odd-valued γs for Seq-SPU did not lose dramatic power due to the sequential algorithm summing the powered score statistics in the correct way. This was most evident when four out of 11 interactions had effect. In this case, all odd-valued γs for Seq-SPU did significantly better than their SPU counterparts. As a result, Seq-aSPU was much better than aSPU in this scenario.

4.4 Minnesota Center for Twin and Family Research

Finally, we applied the methods studied in Section 4.3 to the MCTFR dataset. The MCTFR follows MZ and DZ twins through adolescence into at least early adulthood to study psychological outcomes such as SUDs (Miller et al., 2012). Previous studies of the MCTFR data have estimated the amount of alcohol consumption to be highly heritable with about 50% to 80% of the phenotype explained by genetic variation in the four-member population consisting of parents and genetic offspring (McGue et al., 2013). The development of this SUD reflects the influence of genes that are modulated by environmental factors such as the quality of the parent-child relationship, affiliation with deviant peers, personality characteristics, antisociality (e.g., conduct disorder) and life stress. The goal of this analysis is to perform gene-based tests of GxE interaction to study whether association between drinking score and candidate genes for alcoholism are modified at age 17 in the twin cohort by four different environmental factors.

The phenotype used in our models was a drinking index formed by a factor analysis of questions from an in-person interview and questionnaire. Our environmental factors of interest, 'deviant peers', 'environmental assets', 'family conflict', 'family climate'

scores, were computed as factor scores from a range of questionnaires completed by the twins and their parents. The environmental factor scores are normally distributed and correlated with each other as described in Section 4.3.

We tested for GxE interaction between the SNPs in the candidate genes for alcoholism listed in Olfson and Bierut (2012) and the four environments. 50 of these 54 genes had genotyped SNPs in the MCTFR data. Covariates used in the model include age and the first four principal components from an Eigenstrat analysis (Price et al., 2006) of the SNP data were used as covariates to adjust for population stratification. We used the Joint approach with the ridge penalized ACE model to test for GxE interaction in these genes because this model and approach performed best in our simulations. We limited our analyses to the seventeen-year-olds with non-zero drinking scores to avoid confounding the mechanisms that influence teenagers to start drinking with the mechanisms that influence how much a teenager drinks. Lastly, we analyzed males and females separately because two out of the four environments had a significant interaction with sex. This finding is consistent with previous findings that the disease etiology of alcoholism differs among sex [Prescott, 2002; Hardie et al., 2008]. We also found that the A,C, and E variance components shown in Table 4.3 varied greatly between males and females.

To compute p-values for aSPU and Seq-aSPU, we initially sampled the null score vector with $B = 1000$. For smaller p-values, we increased B by a factor of ten until the final p-value was greater than $5/B$. Figure 4.2 shows p-values for Seq-aSPU compared to the p-values of aSPU for each sex's fifty genes. Only one of these p-values, Seq-aSPU's p-value for CNR1 in the female analysis, met the $\alpha = 0.05/50 = 0.001$ threshold to be identified as a significant GxE interaction. We also found that the largest estimated percent of variance in alcohol consumption explained by a gene was 0.47% for the GABRB1 gene in females and 0.46% for the HTR1B gene in males.

While most of the p-values are very similar for aSPU and Seq-aSPU, in Table 4.3 we investigated the genes, SLC6A2, CNR1, and ADH7, for which the two tests show large differences. In these three genes, the main difference between the SPU and Seq-SPU tests are for $\gamma = 1$. For SLC6A2 and CNR1, iSeq-aSum had a much smaller p-value than the Sum test, which caused Seq-aSPU to have a much smaller p-value than aSPU. Conversely, the Sum test had a much smaller p-value than iSeq-aSum for ADH7

which caused aSPU to have a much smaller p-value than Seq-aSPU. These findings are consistent with our simulations in Section 4.3.

4.5 Discussion

In this paper, we developed tests of interaction between SNPs in a candidate gene and multiple environments for family data using an LMM framework. This framework allows us to incorporate covariance terms to adjust for genetic similarity and unmeasured shared environments in families using the ACE model. We incorporated the measured environments using the MinP and Joint approaches. We found that even if only one environment has interaction with the gene, the Joint approach can still have an advantage over the MinP approach because the environments are correlated. Thus, if there are multiple correlated environments of interest, it is best to test for GxE interaction for the entire set rather than independently. We also proposed using a ridge penalty re-expressed as a random effect for the genetic main effect, which can drastically reduce the number of parameters we need to estimate. By expressing this penalty term as a random effect, we avoid the computational difficulties of penalized mixed models. Finally, we proposed a generalization of the iSeq-aSum test which is equivalent to a weighted version of the SPU test with weights calculated using the sequential algorithm of Basu and Pan (2011). Because aSPU and Seq-aSPU adaptively choose the best γ over the set Γ, if a γ-value included in the search performs poorly, the adaptive test can potentially lose power. For instance, when there were a mix of positive and negative interactions, iSeq-aSum had much higher power than the Sum test; thus, Seq-aSPU was more powerful than aSPU. Thus, by improving the power of odd-valued γs when there are a combination of positive and negative effects, specifically $\gamma = 1$, Seq-aSPU can gain substantial power over aSPU. However, if the interactions were in the same direction, aSPU was more powerful than Seq-aSPU, because the Sum test was more powerful than iSeq-aSum. For our analysis of the MCTFR data, most of the p-values for aSPU and Seq-aSPU were very similar. However, the only significant GxE interaction was identified by Seq-aSPU and not aSPU because iSeq-aSum's p-value was very small while the Sum test was not close to significant.

The models presented here do have some limitations. First, Seq-aSPU is more

computationally intensive than aSPU because it searches for the best way to sum the weighted SPU test for every γ. However, this search only provided a benefit for odd-valued γs in our simulations. Thus, a more computationally efficient approach would be to not perform this search for even γs and instead either use the equivalent SPU test or let $\boldsymbol{a}_\gamma = 1$. Secondly, using LMMs to compute the score vector $\mathbf{U}$ and its covariance $\mathbf{V}$ can also be quite computationally intensive. For our simulations using the `R` package `regress`, this calculation accounted for the majority of our computation time (Appendix B.3). While `R` packages such as `lme4` and `nlme` can be quite fast, these packages are currently not able to implement kinship matrices or a ridge penalty expressed as a random effect. Alternatively, it is possible to calculate the score vector for GxE interaction using generalized estimating equations (GEEs). However, implementing a ridge penalty for the genetic main effect in GEEs is not straightforward. Finally, the models presented here are not intended for binary outcomes. Generalized LMMs with a logit link are much more difficult to fit and ridge penalization can no longer be easily re-cast as a random effect. Once again, a possible alternative is to use a GEE model without a ridge penalty but more study is needed.

Here, we have only studied a cross-section at age 17 for the MCTFR twins. However, we also have measured alcohol consumption for these twins at ages 14, 20, 24, and 29. We are currently extending these tests of GxE interaction to the longitudinal family data.

Table 4.2: **Simulation 2.** Type I error and power comparison of the ACE model with a ridge penalty implemented or not. Simulations used two causal interactions acting in opposite directions. The interactions explain 2% of the total variance of the simulated phenotype. $\hat{\sigma}^2_G$ is multiplied by 11 so that it can be interpreted as the proportion of variance explained by the 11 SNPs in the simulation.

	R^2_G=	**0.005**		**0.01**		**0.1**	
Ridge Penalty:		**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
Mean (SD)	$11 \times \hat{\sigma}^2_G =$	0.005 (0.006)	-	0.011 (0.01)	-	0.16 (0.032)	-
of Variance	$\hat{\sigma}^2_A =$	0.313 (0.095)	0.313 (0.096)	0.312 (0.095)	0.313 (0.096)	0.313 (0.083)	0.313 (0.083)
Parameter Est.	$\hat{\sigma}^2_C =$	0.126 (0.083)	0.125 (0.084)	0.127 (0.083)	0.126 (0.084)	0.129 (0.074)	0.129 (0.074)
	$\hat{\sigma}^2_E =$	0.364 (0.028)	0.364 (0.028)	0.36 (0.028)	0.36 (0.028)	0.272 (0.022)	0.272 (0.022)
Type I error		**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
$\boldsymbol{\alpha = 0.05}$	**Score**	0.0420	0.0418	0.0416	0.0416	0.0441	0.0427
	aSPU	0.0457	0.0474	0.0449	0.0473	0.0465	0.0479
	Seq-aSPU	0.0438	0.0466	0.0439	0.0470	0.0468	0.0479
$\boldsymbol{\alpha = 0.01}$	**Score**	0.0072	0.0076	0.0074	0.0078	0.0077	0.0081
	aSPU	0.0088	0.0094	0.0088	0.0090	0.0089	0.0094
	Seq-aSPU	0.0096	0.0106	0.0091	0.0102	0.0108	0.0111
Power		**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
$\boldsymbol{\alpha = 0.05}$	**Score**	0.6170	0.5950	0.6220	0.6030	0.7370	0.7220
	aSPU	0.7360	0.7350	0.7440	0.7400	0.8230	0.8250
	Seq-aSPU	0.7650	0.7620	0.7620	0.7630	0.8540	0.8480
$\boldsymbol{\alpha = 0.01}$	**Score**	0.4340	0.4040	0.4360	0.4140	0.5640	0.5450
	aSPU	0.5920	0.5790	0.5930	0.5810	0.6920	0.6930
	Seq-aSPU	0.6080	0.5950	0.6050	0.5960	0.7070	0.7110

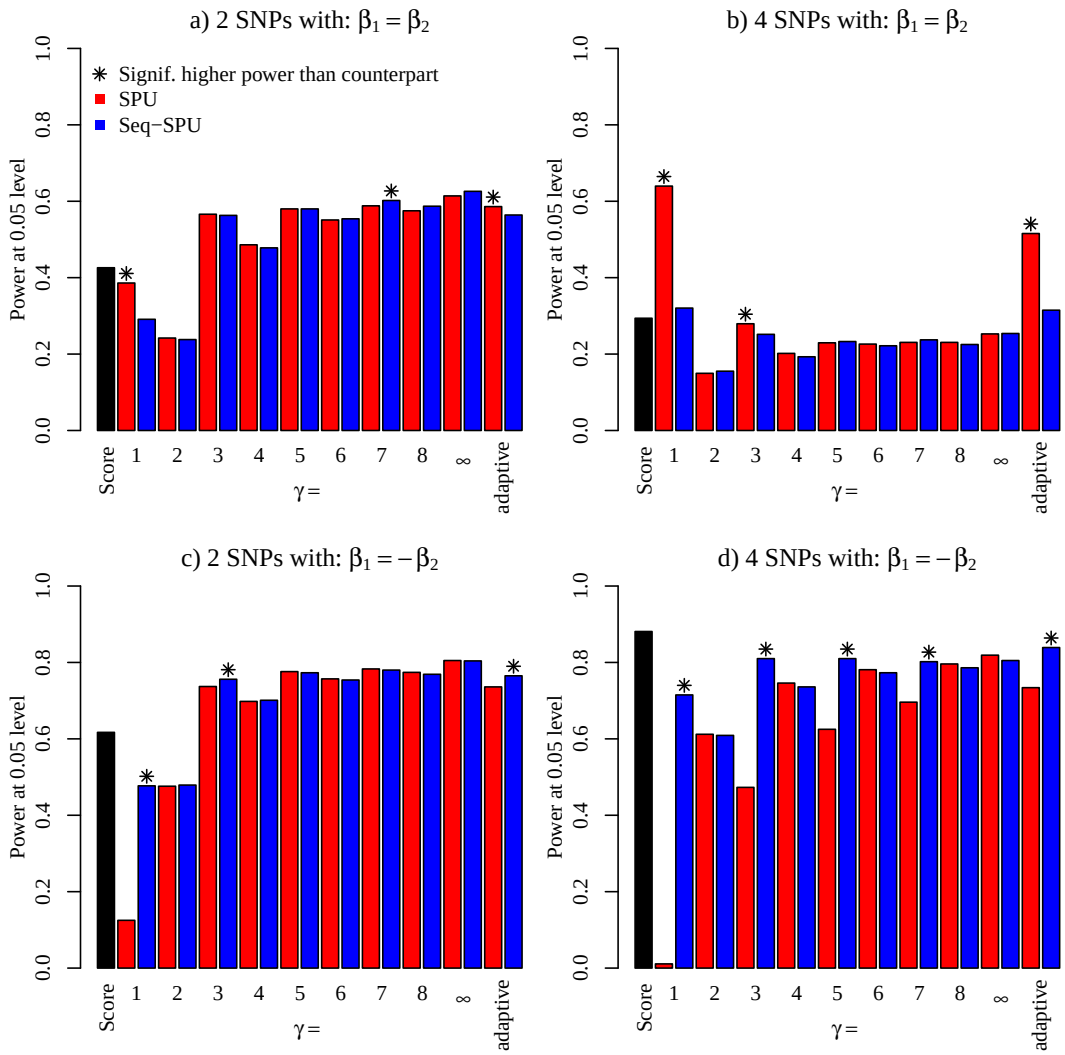

Figure 4.1: **Simulation 3.** Comparison of the SPU and Seq-SPU family of tests in two scenarios: (a,b) interactions in the same direction, OR (c,d) interactions in opposite directions. There are either 2/11 interactions with effect (a,c) OR 4/11 interactions with effect (b,d) that explain 2% of the total variation. For a given γ, significant differences in power between the SPU test and Seq-SPU test are denoted with a star. Significant differences are defined as 95% confidence interval (CI) of the power of Seq-SPU ($power \pm z_{0.025}\sqrt{(0.95)(0.05)/1000}$) not containing the power of its SPU counterpart.

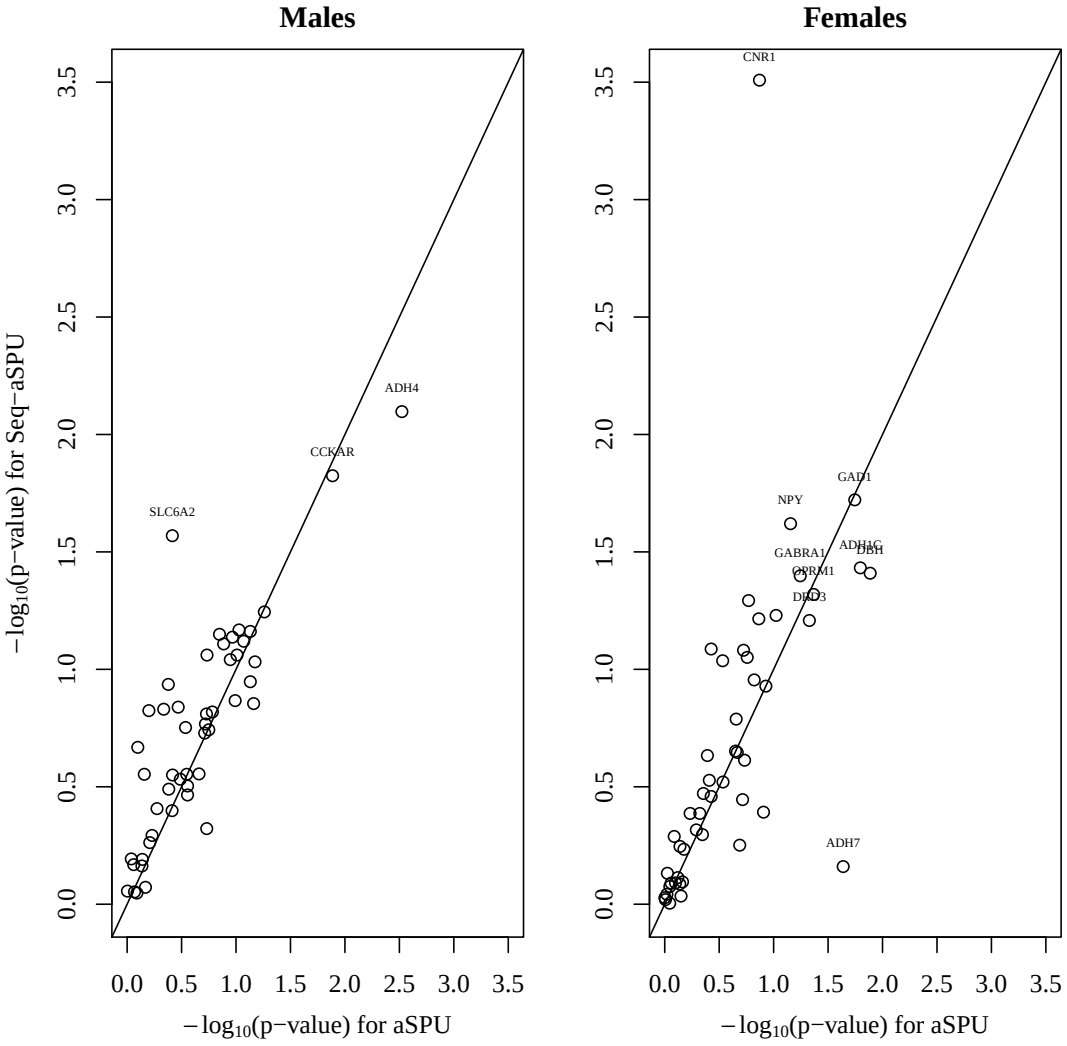

Figure 4.2: **MCTFR analysis.** The -log10(p-values) for aSPU and Seq-aSPU are plotted against each other for our stratified analysis. Points near the diagonal line are genes where aSPU and Seq-aSPU performed similarly. Genes with at least one p-value less than 0.05 are labeled.

Table 4.3: Real data analysis of MCTFR twin data stratified by sex. Genes with large differences between aSPU and Seq-aSPU are shown. P-values less than 0.05 are marked in bold.

	Males		**Females**			
Gene	**SLC6A2**		**ADH7**		**CNR1**	
Location	16q12.2		4q23		6q15	
# SNPs	44		38		39	
# SNPs $\times\hat{\sigma}_G^2 =$	0.0003		0.0022		2.1e-09	
$\hat{\sigma}_A^2 =$	0.410		0.171		0.172	
$\hat{\sigma}_C^2 =$	6.6e-06		0.085		0.088	
$\hat{\sigma}_E^2 =$	0.293		0.273		0.274	
$\gamma =$	**SPU**	**Seq-SPU**	**SPU**	**Seq-SPU**	**SPU**	**Seq-SPU**
1	0.92208	**0.01299**	**0.00899**	0.46653	0.10190	**0.00009**
2	0.17982	0.34466	0.45055	0.59241	**0.04995**	0.09855
3	0.64635	0.32368	0.16384	0.52647	0.07692	0.09337
4	0.32567	0.42358	0.54146	0.62837	0.12887	0.17647
5	0.73127	0.47353	0.33167	0.62438	0.13387	0.19106
6	0.44855	0.51249	0.61039	0.66533	0.21079	0.24257
7	0.80619	0.59441	0.45554	0.65834	0.20480	0.23206
8	0.53646	0.58741	0.65534	0.67433	0.26374	0.27757
∞	0.71728	0.74925	0.71129	0.71429	0.23676	0.23322
adaptive	0.38462	**0.02697**	**0.02298**	0.69131	0.13487	**0.00031**
Score	0.13947		0.82886		0.05701	

Chapter 5

Conclusion and Discussion

5.1 Methodological advances

The work presented in this dissertation showcases several methodological advances in modeling genetic data. The strategy to model the genetic data was to use the sequential algorithm that was initially proposed by Basu and Pan (2011) to test for genetic main effects for a set of SNPs and/or RVs. This algorithm searches potential allocations that best sum the regression estimates of the variants in order to produce a more powerful test without searching over all possibilities. This sequential search can best be viewed as a stepwise model selection procedure. In Chapter 2, we extended this model selection approach to a model-averaging approach for testing for genetic main effect among a group of SNPs and RVs from a candidate gene or region. In Chapters 3 and 4, we extended the algorithm of Basu and Pan (2011) to test for GxE interaction in a candidate gene for either unrelated subjects or families.

We began our work in Chapter 2 where we extended the model selection test for genetic main effect in a candidate gene proposed by Basu and Pan (2011) to a model-averaging test. While model-averaging is typically used in the prediction literature, we explored its usefulness for inference. Model-averaging addresses the potential effect of ordering of the variants on the power of the sequential search used in Basu and Pan (2011). Model selection only selects one path to proceed through, so if two path choices are close at a given step, it may choose the wrong path. By exploring both paths in this situation, we found that model-averaging reduced the order dependency of the model

selection scheme.

In Chapter 3, we extended the sequential algorithm for adaptively summing the genetic main effect in Basu and Pan (2011) to instead test for gene-environment interactions. We proposed doing this using two different models aimed at either creating a summary measure for the main effect to be used for interaction testing (iSeq-aSum-G) or a summary measure for the interaction itself (iSeq-aSum-I). Main-effect-based summary measure tests are potentially a useful tool for studying GxE interaction because they use fewer degrees-of-freedom than testing on the interaction itself. However, Lin et al. (2015) proved that these tests have inflated Type I error if genes and environments are correlated. Nevertheless, we showed in our simulations that under realistic scenarios iSeq-aSum-G and the burden test maintained the appropriate Type I error rate. However, these tests can lose significant power if the GxE interactions cannot be captured using a main effect model. In this situation, using a summary-based measure on the interaction with iSeq-aSum-I performed very well if there were many interactions with effects. By taking the minimum p-value among the two, iSeq-aSum-min was very powerful to test for GxE interaction in all of the situations we studied because iSeq-aSum-G and iSeq-aSum-I each performed well in different situations.

Finally, we extended the methods to test for GxE interaction for studies with unrelated subjects as presented in Chapter 3 to family data in Chapter 4. We used an LMM framework to account for within family correlation by splitting the covariance within a family into three parts: shared genetic effects, unmeasured shared environmental effects, and unshared effects. We also implemented a ridge penalty on the genetic main effect in order to avoid computing a large number of parameters as suggested by Lin et al. (2013) and Lin et al. (2015). However, penalizing a linear mixed model can be very computationally challenging. Instead, we re-expressed this ridge penalty as a random effect that is easy to compute within our mixed model framework. Using the score vector of GxE interaction and its covariance from the resulting LMM, most tests from Chapter 3 could be implemented to test for interaction among a group of SNPs from a candidate gene and a set of environmental factors for family studies. We also proposed a generalization to iSeq-aSum-I which optimizes over different powers of the summation of the score vector for GxE interaction rather than the score vector itself. This generalization is equivalent to a weighted version of the aSPU test (Pan et al., 2014; Kim et al., 2014)

with weights chosen from the sequential algorithm of Basu and Pan (2011). Unlike the odd-powered SPU tests, the odd-powers of our test, Seq-SPU, are able to adaptively sum the positive and negative effects, which can make these tests more powerful than their SPU counterparts. Consequently, the adaptive version of our test, Seq-aSPU, was more powerful than aSPU in these situations. This advantage was highlighted in our analysis of the MCTFR twin data. Because iSeq-aSum-I had a very small p-value while its SPU counterpart did not, Seq-aSPU was the only method to identify the lone significant gene in the study that interacts with the four environmental factors.

5.2 Future directions for research

The work of this dissertation can be extended in many different ways. While the model-averaging test in Chapter 2 reduced the order dependency of the sequential search, this new test was much more computationally intensive and only resulted in a slight power gain over model selection. One possible explanation for this modest power gain could be that the model selection approach already implements a dimension reduction strategy, which requires estimation of only three parameters for each model. Due to the small number of parameters, the uncertainty in model selection decreases and the advantage of model-averaging over model selection becomes less significant. An alternative formulation where we allow the model-averaging approach to have distinct parameters for each directional effect while undergoing variable selection could allow for larger gains in power over model selection.

Chapter 4 was an extension of the methods from Chapter 3 to family data. The advantages shown in Chapter 4 by the Seq-aSPU test over the aSPU test in the presence of positive and negative effects most likely extend to testing problems outside of GxE interaction such as genetic main effect testing of SNPs/RVs and gene-gene interactions. A possible further extension of the SPU and/or Seq-SPU tests for main effect or interactions is to apply these tests to the full longitudinal MCTFR twin sample rather than at only age 17. Longitudinal studies have numerous advantages over cross-sectional studies. With the same sample size as cross-sectional studies, the repeated measurements on individuals could provide better power to detect these genes with small effect sizes. Additionally, between-subject variation is excluded from error and hence might

provide more efficient estimates of the parameters. However, if we wish to test for GxE interaction over time, the addition of time increases the difficulty of our models because GxE interaction effects may change over time.

References

Asimit, J., Day-Williams, A., Morris, A., and Zeggini, E. (2012). ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Hum. Hered.* **73,** 84–94.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. (2011). An application and empirical comparison of statistical analysis methods for associating rare variants to a complex phenotype. *Pacific Symposium on Biocomputing Proceedings* page 16.

Basu, S. and Pan, W. (2011). Comparison of statistical tests for association with rare variants. *Genetic Epidemiology* **35,** 606–619.

Basu, S., Pan, W., and Oetting, W. S. (2011). A dimension reduction approach for modeling multilocus interaction in case-control studies. *Human Heredity* **71,** 234–245.

Bhatia, G., Bansal, V., Harismendy, O., Schork, N., Topol, E., Frazer, K., and Bafna, V. (2010). A covering method for detecting genetic associations between rare variants and common phenotypes. *PLOS Computational Biology* page DOI: 10.1371/journal.pcbi.1000954.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference*, pages 20–23. Berlin : Springer, second edition.

Bůžková, P., Lumley, T., and Rice, K. (2011). Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Annals of Human Genetics* **75,** 36–45.

Chen, H., Meigs, J., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.* **37,** 196–204.

Coombes, B., Basu, S., and McGue, M. (2016). A combination test for detection of gene-environment interaction in cohort studies.

Dai, J., Kooperberg, C., Leblanc, M., and Prentice, R. (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* **99,** 929–44.

Davies, R. (1980). The distribution of a linear combination of chi-square random variables. *J R Stat Soc Ser C* **29,** 323–333.

Dawber, T., Meadors, G., and Moore, F. (1951). Epidemiological approaches to heart disease: the framingham study. *Am. J. Public Health* **41,** 279–286.

Derkach, A., Lawless, J., and Sun, L. (2013). Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. *Genet. Epidemiol.* **37,** 110–121.

Elkins, I., McGue, M., Iacano, W., and Tellegen, A. (1997). Genetic and environmental influences on parent-son relationships: Evidence for increasing genetic influence during adolescence. *Developmental Psychology* **33,** 351–363.

Falconer, D. and Mackay, T. (1981). *Introduction to quantitative genetics*. Longman, New York.

Fan, Q., Verhoeven, V. J., Wojciechowski, R., Barathi, V. A., Hysi, P. G., Guggenheim, J. A., Höhn, R., Vitart, V., Khawaja, A. P., Yamashiro, K., et al. (2016). Meta-analysis of gene-environment-wide association scans accounting for education level identifies additional loci for refractive error. *Nature communications* **7,**.

Galton, F. (1874). On men of science, their nature and their nurture. *Proceedings of the Royal Institution of Great Britain* **7,** 227–236.

Gauderman, W., Zhang, P., Morrison, J., and Lewinger, J. (2013). Finding novel genes by testing g x e interactions in a genome-wide association study. *Genet. Epidemiol.* **37,** 603–13.

Hardie, T. L., Moss, H. B., and Lynch, K. G. (2008). Sex differences in the heritability of alcohol problems. *American Journal on Addictions* **17,** 319–327.

Hicks, B., Schalet, B., Malone, S., Iacano, W., and McGue, M. (2011). Psychometric and genetic architecture of substance use disorder and behavioral disinhibition measures for gene association studies. *Behav. Genet.* **41,** 459–475.

Hicks, B., South, S., DiRago, A., Iacano, W., and McGue, M. (2009). Environmental adversity increases genetic risk for externalizing disorders. *Arch Gen Psychiatry* **66,** 640–648.

Higgins, M., Province, M., Heiss, G., Eckfeldt, J., Ellison, R. C., Folsom, A. R., Rao, D., Sprafka, J. M., and Williams, R. (1996). NHLBI family heart study: objectives and design. *American journal of epidemiology* **143,** 1219–1228.

Hodges, J. S. (2013). *Richly parameterized linear models: additive, time series, and spatial models using random effects.* CRC Press.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science* **14,** 382–401.

Hoffmann, T., Marini, N., , and Witte, J. (2010). Comprehensive approach to analyzing rare genetic variants. *PLoS One* **5,** e13584.

Hsu, L., Jiao, S., Dai, J., Hutter, C., Peters, U., and Kooperberg, C. a. (2012). Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet. Epidemiol.* **36,** 183–94.

Hunter, D. (2005). Gene-environment interactions in human disease. *Nature Review Genetics* **6,** 287–98.

Hutter, C., Mechanic, L., Chatterjee, N., Kraft, P., and Gillanders, E. (2013). Gene-environment interactions in cancer epidemiology: a national cancer institute think tank report. *Genet. Epidemiol.* **37,** 643–57.

Ionita-Laza, I., Buxbaum, J., Laird, N., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics* **7,** e1001289.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Gen.* **92,** 841–853.

Jiao, S., Hsu, L., Bzieau, S., Brenner, H., Chan, A., Chang-Claude, J., Marchand, L., Lemire, M., Newcomb, P., Slattery, M., and Peters, U. (2013). SBERIA: set-based gene-environment interaction test for rare and common variants in complex diseases. *Genet. Epidemiol.* **37,** 452–64.

Kaprio, J. (2012). Twins and the mystery of missing heritability: the contribution of gene–environment interactions. *Journal of internal medicine* **272,** 440–448.

Kim, J., Wozniak, J. R., Mueller, B. A., Shen, X., and Pan, W. (2014). Comparison of statistical tests for group differences in brain functional networks. *NeuroImage* **101,** 681–694.

Ko, Y., Saha-Chaudhuri, P., Park, S., Vokonas, P., and Mukherjee, B. (2013). Novel likelihood ratio tests for screening gene-gene and gene-environment interactions with unbalanced repeated-measures data. *Genetic Epidemiology* **37,** 581–91.

Kooperberg, C. and Leblanc, M. (2008). Increasing the power of identifying gene gene interactions in genome-wide association studies. *Genet. Epidemiol.* **32,** 255–63.

Lee, S., Abecasis, G., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Gen.* **95,** 5–23.

Lee, S., Wray, N., Goddard, M., and Visscher, P. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics* **88,** 294–305.

Lee, S., Wu, M., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13,** 762–775.

Lesnick, T., Papapetropoulos, S., Mash, D., Ffrench-Mullen, J., Shehadeh, L., de Andrade, M., Henley, J., Rocca, W., Ahlskog, J., and Maraganore, D. (2007). A genomic pathway approach to a complex disease: axon guidance and parkinson disease. *PLoS Genetics* **3,** E98.

Lewontin, R. (1964). The interaction of selection and linkage. i. general considerations heterotic models. *Genetics* **49,** 49–67.

Li, B. and Leal, S. (2009). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83,** 311–321.

Lin, D. and Tang, Z. (2012). A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89,** 354–367.

Lin, X., Lee, S., Christiani, D., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **14,** 667–81.

Lin, X., Lee, S., Wu, M., Wang, C., Chen, H., Li, Z., and Lin, X. (2015). Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **DOI: 10.1111/biom.12368,** 1–9.

Liu, D. and Leal, S. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genetics* **6,** 1–14.

Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *JASA* **89,** 1535–1546.

Madsen, B. and Browning, S. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* **5,** e1000384.

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D., McCarthy, M., and Ramos, E. (2009). Finding the missing heritability of complex diseases. *Nature* **461,** 747–753.

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine* **363,** 166–176.

McGue, M., Irons, D., and Iacono, W. G. (2014). The adolescent origins of substance use disorders: A behavioral genetic perspective. In *Genes and the Motivation to Use Substances*, pages 31–50. Springer.

McGue, M., Zhang, Y., Miller, M., Basu, S., Vrieze, S., Hicks, B., Malone, S., Oetting, W., and Iacano, W. (2013). A genome-wide association study of behavioral disinhibition. *Behav. Genet.* **43,** 363–373.

Miller, M., Basu, S., Cunningham, J., Eskin, E., Malone, S., Oetting, W., Schork, N., Sul, J., Iacano, W., and McGue, M. (2012). The Minnesota Center for Twin and Family Research genome-wide association study. *Twin Research and Human Genetics* **15,** 767–774.

Morgenthaler, S. and Thilly, W. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research* **615,** 28–56.

Mukherjee, B., Ko, Y., VanderWeele, T., Roy, A., Park, S., and Chen, J. (2011). Principal interactions analysis for repeated measures data: application to gene-gene and gene-environment interactions. *Statistics in Medicine* **31,** 2531–51.

Murcray, C., Lewinger, J., Conti, D., Thomas, D., and Gauderman, W. (2011). Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet. Epidemiol.* **35,** 201–10.

Murcray, C., Lewinger, J., and Gauderman, W. (2009). Gene-environment interaction in genome-wide association studies. *Am. J. Hum. Gen.* **169,** 219–26.

Neale, B., Rivas, M., Voight, B., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S., Roeder, K., and Daly, M. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics* **7,** e1001322.

Olfson, E. and Bierut, L. (2012). Convergence of GWA and candidate gene studies for alcoholism. *Alcohol Clin Exp Res* **36,** 2086–94.

Ottman, R. (1990). An epidemiologic approach to gene-environment interaction. *Genetic Epidemiology* **7,** 177. doi:10.1002/gepi.1370070302.

Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology* **33,** 497–507.

Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics* **197,** 1081–95.

Pan, W., Shen, X., and Basu, S. (2011). Adaptive tests for detecting gene-gene and gene-environment interactions. *Hum. Hered.* **72,** 98–109.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J., Jin, L., Amos, C., and Xiong, M. (2009). Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal of Human Genetics* **18,** 111–17.

Pinheiro, J. and Chao, E. (2006). Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics* **15,** 58–81.

Prescott, C. A. (2002). Sex differences in the genetic risk for alcoholism. *Alcohol Research and Health* **26,** 264–273.

Price, A., Kryukov, G., de Bakker, P., Purcell, S., Staples, J., LJ, W., and Sunyaev, S. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics* **86,** 982.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38,** 904–909.

Raftery, A., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *JASA* **92,** 179–191.

Samek, D. R., Hicks, B. M., Keyes, M. A., Iacono, W. G., and McGue, M. (2016). Antisocial peer affiliation and externalizing disorders: Evidence for gene x environment x development interaction. *Development and Psychopathology* **FirstView,** 1–18.

Schork, N., Murray, S., Frazer, K., and Topol, E. (2009). Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development* **19,** 212–219.

Shen, X., Alam, M., Fikse, F., and Ronnegard, L. (2013). A novel generalized ridge regression method for quantitative genetics. *Genetics* **193,** 1255–1268.

Simonds, N. I., Ghazarian, A. A., Pimentel, C. B., Schully, S. D., Ellison, G. L., Gillanders, E. M., and Mechanic, L. E. (2016). Review of the gene-environment interaction literature in cancer: What do we know? *Genetic Epidemiology* **40,** 356–365.

Slatkin, M. (2009). Epigenetic inheritance and the missing heritability problem. *Genetics* **182,** 845–850.

Speed, D. and Balding, D. J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome research* **24,** 1550–1557.

Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genet. Epidemiol.* **37,** 334–344.

Thomas, D. (2010a). Gene-environment-wide association studies: Emerging approaches. *Nature Review Genetics* **11,** 259–272.

Thomas, D. (2010b). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu Rev Public Health* **31,** 21–36.

Viallefont, V., Raftery, A., and Richardson, S. (2001). Variable selection and Bayesian model averaging in epidemiological case-control studies. *Statistics in Medicine* **20,** 3215–3230.

Visvikis-Siest, S. and Siest, G. (2008). The STANISLAS cohort: a 10-year follow-up of supposed healthy families. gene-environment interactions, reference values and evaluation of biomarkers in prevention of cardiovascular diseases. *Clinical chemistry and laboratory medicine* **46,** 733–747.

Vrieze, S., Feng, S., Miller, M., Hicks, B., Pankratz, N., Abecasis, G., Iacano, W., and McGue, M. (2014). Rare nonsynonymous exonic variants in addiction and behavioral disinhibition. *Biological Psychiatry* **75,** 783–789.

Wang, Y., Li, D., and Wei, P. (2015). Powerful Tukey's one degree-of-freedom test for detecting gene–gene and gene–environment interactions. *Cancer Informatics* **14,** 209–18.

Wray, N. and Visscher, P. (2008). Estimating trait heritability. *Nature Education* **1,** 29.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am. J. Hum. Genet.* **89,** 82–93.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88,** 76–82.

Yang, Q. and Khoury, M. (1997). Evolving methods in genetic epidemiology. iii. gene-environment interaction in epidemiologic research. *Epidemiologic Reviews* **19,** 33–43.

Yu, K., Wacholder, S., Wheeler, W., Wang, Z., Caporaso, N., Landi, M., and Liang, F. (2012). A flexible bayesian model for studying gene–environment interaction. *PLoS Genet* **8,** e1002482.

Zhang, Z., Sha, Q., Wang, X., and Zhang, S. (2011). Detection of rare variant effects in association studies: extreme values, iterative regression, and a hybrid approach. *BMC Proceedings* **5,** S112.

Zuk, O., Hechter, E., Sunyaev, S., and Lander, E. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109,** 1193–1198.

Appendix A

Supplementary materials for Chapter 3

A.1 Proof of asymptotic distribution of iSeq-aSum-I

Proposition:- The statistic {iSeq-aSum}-I in Section 3.2.3 will have a χ^2 distribution with approximately $\frac{3Q}{4}$ degrees of freedom for a large Q number of SNPs under the null hypothesis of no gene-environment interaction.

Proof:

Assume we have n unrelated individuals with Q common genetic variants from a candidate gene, K measured covariates, and an environmental factor. Let Y_i, E_i, $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{iK})$ be the phenotype, environmental factor, and K covariates for the i^{th} individual, respectively. Let $\mathbf{G}_i = (G_{i1}, \cdots, G_{iQ})$ be the genotypes for the Q variants each standardized by their mean and standard deviation. Define $\mathbf{S}_i = (G_{i1}E_i, \cdots, G_{iQ}E_i) = (S_{i1}, \cdots, S_{iQ})$ to be the Q pairwise G-E interactions for the i^{th} individual. We will assume that our phenotype is continuous although all of these methods can be extended to binary phenotypes. Let $\boldsymbol{Y}^{'} = (Y_1^{'}, \cdots, Y_n^{'})$ be the residuals from the following main-effect model:

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\alpha_{1,k} + E_i\alpha_2 + \sum_{q=1}^{Q} G_{iq}\alpha_{3,q} + \epsilon_i^{'}; \quad i = 1, ..., n \tag{A.1}$$

where $\epsilon_i^{'} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2)$ and α_0, $\boldsymbol{\alpha}_1 = (\alpha_{1,1}, \cdots, \alpha_{1,K})$, α_2, $\boldsymbol{\alpha}_3 = (\alpha_{3,1}, \cdots, \alpha_{3,Q})$ are the

effect estimates for the intercept, covariates, environment and SNPs, respectively. To implement the {iSeq-aSum}-I algorithm in Section 2.4.2, we work with these residuals $\boldsymbol{Y}^{'}$.

For Q SNPs, the stepwise model selection algorithm in iSeq-aSum-I as described in Section 3.2.3 browses through $2Q+1$ models. In other words, the optimal allocations for $(\gamma_1, \gamma_2, \ldots, \gamma_Q)$ in Equation 3.5 is selected by browsing through $2Q+1$ possible allocations. Under the null hypothesis, it is equally likely for the data to be best supported by any of these $2Q+1$ allocations. For every allocation, our problem of maximizing the likelihood can be viewed as a clustering problem in the covariate space. The whole model selection scheme described for iSeq-aSum-I approach can be viewed as a maximum likelihood estimation problem for the parameters in the following linear regression model:

$$\begin{aligned} f(\boldsymbol{Y}^{'}, \mathbf{G}) &= \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma_e^2} \left(Y_i^{'} - \beta_c \sum_{q=1}^{Q} \gamma_q S_{iq} \right)^2 \right) \quad \gamma_j = -1, 0, 1 \\ &= \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma_e^2} \left(Y_i^{'} - \beta_c \sum_{q^{'} \in Gr1} S_{iq} + \beta_c \sum_{q \in Gr2} S_{iq} \right)^2 \right), \end{aligned}$$

where $Gr1$ represents the positively associated ('+1' group) variants and $Gr2$ represents the negatively associated ('-1' group) variants with the disease. Under the null hypothesis of no association, it is equally likely for a variant to belong to any of these two groups. Hence the null can be achieved either through $\beta_c = 0$ or by assigning equal number of variants into the two groups '$Gr1$' and '$Gr2$' under the assumption that the variants are identically distributed. Hence under the null hypothesis, the above likelihood will be maximum if half of these Q variants belong to $Gr1$ and the rest belongs to $Gr2$.

Now the degrees of freedom of our test statistic in Equation 3.7 will depend on the number of SNPs browsed to achieve success that is the equal distribution of variants in high-risk and low-risk group. For example, consider an association analysis with Q=3 SNPs. Our model selection scheme with 3 SNPs would consider $2Q+1 = 7$ possible allocations. A possible choice of sequential selection path would be { (1,1,1); (-1,1,1); (0,1,1); (-1,-1,1); (-1,0,1); (-1,0,0); (-1,0,-1) }, where (1,1,1) corresponds to the allocation

for all 3 SNPs in the high-risk group.

For a specific set of $(2Q+1)$ allocations as illustrated above, the likelihood ratio test statistic in Equation 3.7 will have a χ^2 distribution with degrees of freedom equal to the number of SNPs browsed to achieve equal number of SNPs in '$Gr1$' and '$Gr2$'. In the example described above, it was achieved in the 5th allocation after browsing the first 2 SNPs. Hence under the null hypothesis T_I would follow chi-square distribution with degrees of freedom as the # of SNPs (z_Q, say) browsed to achieve equal representation in two groups. Without loss of generality, let us assume that Q is even. Note that the algorithm needs to browse through at least $Q/2$ allocations to achieve equal representation of SNPs in '$Gr1$' and '$Gr2$'. Under the null hypothesis, z_Q will be uniformly distributed between $Q/2$ and Q. Note that under the null hypothesis

$$T_I \sim \chi^2_{z_Q} z_Q \Rightarrow E(z_Q) = \frac{1}{Q}[Q/2 + \ldots + Q] = 3Q/4 \tag{A.2}$$

A.2 Correlation between burden summary measure and environment

The dependent environment is simulated from the Equation 3.9. Allow the first four SNPs for the burden summary measure to be the causal SNPs that are correlated with the environment. The correlation between the burden summary measure $\sum_{q=1}^{Q} G_q$ and the environment E:

$$\text{cor}\left(\sum_{q=1}^{Q} G_q, E\right) = \frac{\text{cov}\left(\sum_{q=1}^{Q} G_q, E\right)}{\sqrt{\text{var}\left(\sum_{q=1}^{Q} \mathbf{G}_q\right)}\sqrt{\text{var}(\mathbf{E})}}$$

where for uncorrelated SNPs

$$\text{cov}\left(\sum_{q=1}^{Q} G_q, E\right) = \text{cov}\left(\sum_{q=1}^{Q} G_q, \phi\sum_{q=1}^{4} G_q + \epsilon\right) = \phi\,\text{var}\left(\sum_{q=1}^{4} \mathbf{G}_q\right)$$

$$\text{var}(\mathbf{E}) = \text{var}\left(\phi\sum_{q=1}^{4} G_q + \epsilon\right) = \phi^2\,\text{var}\left(\sum_{q=1}^{4} G_q\right) + 1$$

$$\text{var}\left(\sum_{q=1}^{Q} G_q\right) = \text{var}\left(\sum_{q=1}^{4} G_q\right) = \sum_{q=1}^{4} \text{var}(G_q).$$

For uncorrelated and standardized SNPs, the correlation will be $\frac{4\phi}{2\sqrt{4\phi^2+1}}$. Since the SNPs in ADH1B and ALDH1A1 were correlated, we used the parent population of MCTFR cohort to estimate the correlation between the burden summary measure and the environment for Figure 2.

A.3 Performance of minP test for interaction testing

Proof: The p-value for the minP test is calculated as $Q \times \min_{q=1,\cdots,Q}\{p_q\}$ where each p_q p-value is calculated separately for the q^{th} interaction from Model 3.2. Let the k^{th} p-value for the k^{th} interaction model be the minimum. For this model, the t-test statistic for the estimated interaction parameter $\hat{\beta}_k^*$ is:

$$T_k = \frac{\hat{\beta}_k^*}{SE_k}$$

where $SE_k = \frac{\text{RSS}/(n-K-4)}{\sqrt{\text{var}(G_{ik})}}$ is the standard error for β_k where $\text{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is the residual sum of squares. However, if the true model is Model 3.1,

$$\begin{aligned}
\text{RSS} =& \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\
=& \sum_{i=1}^{n}\left(\alpha_0 + \sum_{k=1}^{K} X_{ik}\alpha_{1,k} + E_i\alpha_2 + \sum_{q=1}^{Q} G_{iq}\alpha_{3,q} + \sum_{q=1}^{Q} S_{iq}\beta_q \right. \\
& \left. -\hat{\alpha}_0^* - \sum_{k=1}^{K} X_{ik}\hat{\alpha}_{1,k}^* - E_i\hat{\alpha}_2^* - G_{ik}\hat{\alpha}_{3,k}^* - S_{ik}\hat{\beta}_k^*\right)^2 \\
=& \sum_{i=1}^{n}\left((\alpha_0 - \hat{\alpha}_0^*) + X_{ik}(\alpha_{1,k} - \hat{\alpha}_{1,k}) + E_i(\alpha_2 - \hat{\alpha}_2^*) + \sum_{q\neq k} G_{iq}\alpha_{3,q} \right. \\
& \left. +G_{ik}(\alpha_{3,k} - \hat{\alpha}_{3,k}^*) + \sum_{q\neq k} S_{iq}\beta_q + S_{i,k}(\beta_k - \hat{\beta}_k^*)\right)^2 \\
=& \sum_{i=1}^{n}\left(\widehat{\text{PE}_i} + \sum_{q\neq k} S_{iq}\beta_q + \sum_{q\neq k} G_{iq}\alpha_{3,q}\right)^2
\end{aligned}$$

where $\widehat{\text{PE}_i}$ is the prediction error for the i^{th} individual corresponding to Model 3.2. Therefore, as we increase $\alpha_{3,q}$ for at least one $q \neq k$, the RSS will increase and thus a

decrease in T_k. This results in a loss of power for the t-test of $\hat{\beta}_k^*$.

Appendix B

Supplementary materials for Chapter 4

B.1 Ridge regression as a random effect

Assume we have independent subjects. Suppose we have a general main-effect only linear model we wish to estimate:

$$\boldsymbol{Y} = \tilde{\boldsymbol{X}}\boldsymbol{\alpha}_X + \mathbf{G}\boldsymbol{\alpha}_3 + \boldsymbol{\epsilon} \tag{B.1}$$

where $\tilde{\boldsymbol{X}} = [\mathbf{1}|\boldsymbol{X}|\mathbf{E}]$, $\boldsymbol{\alpha}_X$ contains the regression coefficients for the intercept, covariates, environmental factors, $\boldsymbol{\alpha}_3 \sim \text{MVN}(\mathbf{0}, \sigma_s^2\mathbf{I}_q)$ contains the random effects for the q SNPs, and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2)$. Given $\hat{\sigma}_e^2$ and $\hat{\sigma}_s^2$, it is easy to show via Henderson's mixed model equations that the estimates for $\boldsymbol{\alpha}_X$ and $\boldsymbol{\alpha}_3$ are

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}_X \\ \hat{\boldsymbol{\alpha}}_3 \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^T\mathbf{G} \\ \mathbf{G}^T\tilde{\boldsymbol{X}} & \mathbf{G}^T\mathbf{G} + \hat{\lambda} \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\boldsymbol{X}}^T \\ \mathbf{G}^T \end{pmatrix} \frac{1}{\hat{\sigma}_e^2}\boldsymbol{Y} \tag{B.2}$$

where $\hat{\lambda} = \hat{\sigma}_e^2/\hat{\sigma}_s^2$ is analogous to ridge penalization.

B.2 AE model estimated covariance when the true model is ACE

If we simulate data using the ACE model, the covariance matrix for an MZ or DZ twin pair is:

$$\mathbf{\Sigma}_{MZ} = \begin{bmatrix} \sigma_A^2 + \sigma_C^2 + \sigma_E^2 & \sigma_A^2 + \sigma_C^2 + \sigma_E^2 \\ \sigma_A^2 + \sigma_C^2 + \sigma_E^2 & \sigma_A^2 + \sigma_C^2 + \sigma_E^2 \end{bmatrix}, \mathbf{\Sigma}_{DZ} = \begin{bmatrix} \sigma_A^2 + \sigma_C^2 + \sigma_E^2 & 1/2\sigma_A^2 + \sigma_C^2 + \sigma_E^2 \\ 1/2\sigma_A^2 + \sigma_C^2 + \sigma_E^2 & \sigma_A^2 + \sigma_C^2 + \sigma_E^2 \end{bmatrix}$$

According to our simulations, the AE misspecified model estimates $\tilde{\sigma}_A^2 = \sigma_A^2 + \sigma_C^2$ where $\tilde{\sigma}_A^2$ is the approximate variance estimate for the A component of the AE model. This results in a covariance matrix for a DZ twin pair as:

$$\mathbf{\Sigma}_{DZ} = \begin{bmatrix} \sigma_A^2 + \sigma_C^2 + \sigma_E^2 & 1/2\sigma_A^2 + 1/2\sigma_C^2 + \sigma_E^2 \\ 1/2\sigma_A^2 + 1/2\sigma_C^2 + \sigma_E^2 & \sigma_A^2 + \sigma_C^2 + \sigma_E^2 \end{bmatrix}$$

There is no difference for MZ pairs. Thus, the only difference between the ACE model and the misspecified AE model is that the off-diagonal element for DZ pairs is biased by $1/2\sigma_C^2$.

B.3 Computation time

Table B.1: Mean computation time in our simulations. The computation time for either method is computed after **U** and **V** are estimated through our LMM fit. The computation time for the score test is neglible given **U** and **V**.

Fitting the LMM	Time (s)
ACE model w/ ridge penalty	522.9
AE model w/ ridge penalty	395.8
ACE model	212.0
One environment (11 interactions)	
SPU	0.7
Seq-aSPU	3.4
Four environments (44 interactions)	
aSPU	0.7
Seq-aSPU	24.0