

2020

Statistical methods for genetic association studies: detecting gene x environment interaction in rare variant analysis

<https://hdl.handle.net/2144/41993>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**STATISTICAL METHODS FOR GENETIC ASSOCIATION STUDIES:
DETECTING GENE X ENVIRONMENT INTERACTION IN RARE VARIANT
ANALYSIS**

by

ELISE LIM

B.S., Carnegie Mellon University, 2013
M.S., Carnegie Mellon University, 2015

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2020

© 2020 by
ELISE LIM
All rights reserved

Approved by

First Reader

Ching-Ti Liu, Ph.D.
Professor of Biostatistics

Second Reader

Josée Dupuis, Ph.D.
Professor of Biostatistics

Third Reader

L. Adrienne Cupples, Ph.D.
Professor of Biostatistics and of Epidemiology

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my research and thesis advisor Dr. Ching-Ti Liu for his continuous support and guidance throughout my graduate school journey. I'm sure it wasn't easy dealing with a student like me for 5 years so I'm very grateful you didn't abandon me. I could not have asked for a better mentor for my PhD study and I would not have finished my dissertation without him. Thank you so much for all you've done.

I'm also grateful for Dr. Josée Dupuis and Dr. L. Adrienne Cupples for their invaluable advice and constructive feedback on not only my dissertation but also on my research projects. I would also like to thank Dr. Douglas Kiel and Dr. Gina Peloso for providing insightful and helpful review of this dissertation. Special thanks to Dr. Han Chen for always willing to help and review my work.

Thank you to many friends who always allowed me to vent about my PhD life for as long as I need. I would not have survived if it weren't for all the happy distractions you provided to rest my mind. Special thanks to Dorothy for providing endless support and encouragement during my never-ending school life.

I would not have made it this far without the unconditional love, support, and encouragement from my parents. Thank you for always believing in me and inspiring me to follow my dreams.

I also thank my hedgehog Alfredo for giving me tremendous emotional support during my PhD life.

**STATISTICAL METHODS FOR GENETIC ASSOCIATION STUDIES:
DETECTING GENE X ENVIRONMENT INTERACTION IN RARE VARIANT
ANALYSIS**

ELISE LIM

Boston University Graduate School of Arts and Sciences, 2020

Major Professor: Ching-Ti Liu, Professor of Biostatistics

ABSTRACT

Investigators have discovered thousands of genetic variants associated with various traits using genome-wide association studies (GWAS). These discoveries have substantially improved our understanding of the genetic architecture of many complex traits. Despite the striking success, these trait-associated loci collectively explain relatively little of disease risk. Many reasons for this unexplained heritability have been suggested and two understudied components are hypothesized to have an impact in complex disease etiology: rare variants and gene-environment (GE) interactions. Advances in next generation sequencing have offered the opportunity to comprehensively investigate the genetic contribution of rare variants on complex traits. Such diseases are multifactorial, suggesting an interplay of both genetics and environmental factors, but most GWAS have focused on the main effects of genetic variants and disregarded GE interactions. In this dissertation, we develop statistical methods to detect GE interactions for rare variant analysis for various types of outcomes in both independent and related samples. We leverage the joint information across a set of rare variants and implement

variance component score tests to reduce the computational burden. First, we develop a GE interaction test for rare variants for binary and continuous traits in related individuals, which avoids having to restrict to unrelated individuals and thereby retaining more samples. Next, we propose a method to test GE interactions in rare variants for time-to-event outcomes. Rare variant tests for survival outcomes have been underdeveloped, despite their importance in medical studies. We use a shrinkage method to impose a ridge penalty on the genetic main effects to deal with potential multicollinearity. Finally, we compare different types of penalties, such as least absolute shrinkage selection operator and elastic net regularization, to examine the performance of our second method under various simulation scenarios. We illustrate applications of the proposed methods to detect gene x smoking interaction influencing body mass index and time-to-fracture in the Framingham Heart Study. Our proposed methods can be readily applied to a wide range of phenotypes and various genetic epidemiologic studies, thereby providing insight into biological mechanisms of complex diseases, identifying high-penetrance subgroups, and eventually leading to the development of better diagnostics and therapeutic interventions.

TABLE OF CONTENTS

Chapter 1 Introduction	1
1.1 Genetic association studies	1
1.2 Rare variant analysis	4
1.3 Gene-environment interaction	5
1.4 Family data.....	6
1.5 Dichotomous and survival outcomes	7
1.6 Dissertation outline	7
Chapter 2 Methods for Detecting Gene by Environment Interaction in Rare Variant Analysis for Binary and Continuous Traits in Family Data	9
2.1 Introduction.....	9
2.2 Generalized linear mixed model	13
2.3 Interaction test in generalized linear mixed model for family data	13
2.4 Estimation and hypothesis testing.....	15
2.5 Simulation Studies	17
2.5.1 Type I Error.....	17
2.5.1.1 Type I Error Simulation Settings	17
2.5.1.2. Simulation Results for Type I error	19
2.5.2 Power.....	21
2.5.2.1 Power Simulation Settings.....	21
2.5.2.2 Simulation Results for Power of the test of interaction only	22
2.6 Application to the Framingham Heart Study	25

2.7 Discussion.....	28
Chapter 3 Methods for Detecting Gene by Environment Interaction in Rare Variant Analysis for Time-to-Event Outcomes	31
3.1 Introduction.....	31
3.2 Survival analysis.....	33
3.3 Cox proportional hazards model.....	35
3.4 Regularized regression.....	37
3.5 Gene-environment interaction test for survival outcomes.....	38
3.6 Simulation Studies.....	42
3.6.1 Type I Error.....	42
3.6.1.1 Type I Error Simulation Settings.....	42
3.6.1.2 Type I Error Results.....	45
3.6.2 Power.....	49
3.6.2.1 Power Simulation Settings.....	49
3.6.2.2 Power Results.....	50
3.7 Application to the Framingham Heart Osteoporosis Data.....	51
3.8 Discussion.....	54
Chapter 4 Exploring Regularization Methods in Detecting Gene by Environment Interaction in Rare Variant Analysis for Time-to-Event Outcomes.....	58
4.1 Introduction.....	58
4.2 LASSO.....	59
4.2.1 LASSO in coxGE.....	61

4.3 Elastic Net.....	62
4.3.1 Elastic Net in coxGE.....	62
4.4 Simulation.....	64
4.4.1 Type I Error.....	64
4.4.1.1 Type I Error Simulation Settings.....	64
4.4.1.2 Type I Error Results for Interaction Test.....	66
4.4.2 Power.....	70
4.4.2.1 Power Simulation Settings.....	70
4.4.2.2 Power Results.....	71
4.5 Application to the Framingham Heart Osteoporosis Data.....	77
4.6 Discussion.....	79
Chapter 5 Summary and Future Work	83
5.1 Summary.....	83
5.2 Future Work.....	84
5.2.1 Extension of famGE.....	84
5.2.2 Extension of coxGE.....	84
Appendices.....	86
Appendix A: Derivation of the famGE statistic to test interaction.....	86
Appendix B: Testing interaction with genotypes as fixed effects	90
Appendix C: QQ plot of famGE applied genome-wide to detect gene x smoking interaction on BMI in the Framingham Heart Study data	91

Appendix D: Simulation for type 1 errors of coxGE with sample size of 1000 and causal markers all in positive directions	92
Appendix E: Simulation for type 1 errors for coxGE with sample size of 1000 and causal markers in both positive and negative directions.....	93
Appendix F: Power for coxGE and rareGE with 20%, 50%, and 80% causal variants where the directions of the risk alleles for genetic main effects and GE interactions are opposite.....	94
Appendix G: Type 1 error results for coxGE with ridge, LASSO, and elastic net penalties and empirical significance threshold for low LD setting.....	95
Appendix H: Type 1 error results for coxGE with ridge, LASSO, and elastic net penalties and empirical significance threshold for high LD setting	96
Appendix I: Correlation structure for 20% causal variants in the model	97
Appendix J: Correlation structure for 50% causal variants in the model	97
Bibliography.....	98
Curriculum Vitae... ..	111

LIST OF TABLES

Table 1: Comparison of famGE, rareGE, and BT type 1 errors based on 20,000 replications (in %)	20
Table 2: Association results of gene by smoking interaction on continuous trait: BMI and binary trait: overweight status ($BMI \geq 25$)	27
Table 3: Type 1 error results for rareGE and coxGE and empirical significance threshold for coxGE with genotype main effects where the directions of the risk alleles are all positive ($n=3000$)	47
Table 4: Type 1 error results for rareGE and coxGE and empirical significance threshold for coxGE with genotype main effects where the directions of the risk alleles are 50% positive and 50% negative ($n=3000$)	48
Table 5: Association results of gene by smoking interaction on time-to-fracture in the Framingham Osteoporosis Study	54
Table 6: Type 1 error results for interaction using coxGE with ridge, LASSO, and elastic net penalties for low LD setting at asymptotic thresholds $\alpha = 0.01$ and 0.001	68
Table 7: Type 1 error results for interaction using coxGE with ridge, LASSO, and elastic net penalties for high LD setting at asymptotic thresholds $\alpha = 0.01$ and 0.001	69
Table 8: Percentage of causal variants that were kept in the model in high LD setting	77
Table 9: Association results of gene by smoking interaction on time-to-fracture with ridge, LASSO, and elastic net approaches	78

Table 10: Number of variants included in each gene for the association of gene by
smoking interaction on time-to-fracture with ridge, LASSO, and elastic net penalties
..... 79

LIST OF FIGURES

Figure 1: Power comparison of famGE and BT for continuous trait.....	23
Figure 2: Power comparison of famGE and BT for binary trait.....	24
Figure 3: Power for coxGE and rareGE with 20%, 50%, and 80% causal variants where the directions of the risk alleles for genetic main effects and GE interactions are all positive.....	51
Figure 4: Power comparison of coxGE with ridge, LASSO, and elastic net penalties with 20% (A), 50% (B), and 80% (C) causal variants in the model in low LD setting....	73
Figure 5: Power comparison of coxGE with ridge, LASSO, and elastic net penalties with 20% (A), 50% (B), and 80% (C) causal variants in the model in high LD setting ..	75

LIST OF ABBREVIATIONS

BMD	Bone Mineral Density
BMI	Body Mass Index
BT	Burden Test
coxGE	Cox Gene-Environment test
famGE	Family-Based Gene-Environment test
famSKAT	Family-Based Sequence Kernel Association Test
FHS	Framingham Heart Study
GE	Gene-Environment
GLMM	Generalized Linear Mixed Model
GWAS	Genome Wide Association Studies
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium
LMM	Linear Mixed Model
MAF	Minor Allele Frequency
ML	Maximum Likelihood
MONSTER	Minimum p-value Optimized Nuisance parameter Score Test Extended to Relatives
PQL	Penalized Quasi-Likelihood
RC-SKAT	Rare and Common Sequence Kernel Association Test
REML	Restricted Maximum Likelihood
SKAT	Sequence Kernel Association Test

SKAT-O Sequence Kernel Association Test – Optimal Test

SNP Single Nucleotide Polymorphism

Chapter 1 Introduction

1.1 Genetic association studies

Genome-wide association studies (GWAS) are widely used to scan genetic markers across the genome to detect genetic variations, called single nucleotide polymorphisms (SNPs), associated with complex human diseases or traits, such as schizophrenia, cardiovascular disease, and waist circumference. The underlying assumption in performing GWAS is that common diseases and traits are attributable to common variants, typically defined as those with minor allele frequency (MAF) greater than 1 or 5%. Because GWAS examine SNPs across the whole genome, they are promising ways to study complex diseases and traits since many genetic variants contribute to a person's risk. GWAS have been strikingly successful at discovering hundreds of trait-associated genetic variants, thereby helping to further understand the mechanisms and functions underlying complex disease etiology (Artigas et al. 2011; Chasman et al. 2011). Despite the success, significant common variants discovered by GWAS collectively explain only a small fraction of the estimated heritability of disease risk, thus challenging researchers to further identify factors contributing to disease risk unexplained by GWAS findings (de los Campos, Sorensen, and Gianola 2015; Manolio et al. 2009).

Unexplained heritability could be due to limited sample size in single cohort or case-control studies. Sample size is directly related to power to detect an association, so increasing sample size will further lead to discovery of variants that were previously unidentified in GWAS due to lack of statistical power. As such, cohorts studying the

same trait may combine their study results with meta-analysis to boost power, thus leading to discoveries of novel associated variants (Evangelou et al. 2014; Gorski et al. 2017; Hancock et al. 2010). Meta-analyses can greatly increase sample size and statistical power, and thereby reduce the unexplained heritability.

Environmental variables and possible interaction between genetic and environmental factors are thought to have a big impact on the etiology of complex diseases, thereby offering new insights into unexplained heritability and identifying novel associations (Hamza et al. 2011; Matsui and Ehrenreich 2016). It is well-established that the majority of complex diseases are multifactorial, suggesting an interplay of both the genetics and environmental factors; so examining just the main effects from either genetic or environmental variables, cannot provide full insight into the biological mechanisms of complex diseases and traits. Including environmental variables in the model may help to identify previously undetected loci. This suggests that accounting for environmental variables in genetic analyses may facilitate novel gene discovery, highlight novel biological functions, and aid in uncovering novel gene-environment (GE) interactions that may contribute to explaining variability in the phenotype of interest. Investigating GE interactions can elucidate the etiology of complex diseases, identify subgroups that are at high risk, and eventually lead to the development of better diagnostics and targeted prevention methods because modification of environmental factors, such as diet and smoking, is more feasible than modification of our genome (Zhang, Lin, and Biswas 2017; Zhao et al. 2015). Standard GWAS solely focus on the main effects of genetic

variants, so variants that interact with environmental variables may be missed. Many GE interaction methods have been developed, and including interaction terms in the model might help unravel the genetic architecture of complex diseases. (Chen, Meigs, and Dupuis 2014; Lim et al. 2019; Lin et al. 2013; Ma, Clark, and Keinan 2013; Moreno-Macias et al. 2010; Tzeng et al. 2011)

Another potential explanation for unexplained heritability is the presence of low frequency variants, defined as those with MAF between 1 and 5%, and rare variants with MAF less than 1%. There is increasing empirical evidence that implicates rare variants with modest/large effect sizes in various complex diseases, thereby leading to a new hypothesis that rare sequence variants with relatively high penetrance are a contributor to complex disease susceptibility (Lee et al. 2014). As a result, the focus has shifted from investigating common variants in GWAS to association between rare variants and complex diseases. Rapid advances and decreasing cost of whole genome sequencing technology have facilitated the accessibility and discovery of a plethora of low frequency and rare variants, and yet, substantial statistical, analytical, and computational challenges remain in uncovering rare variants in association studies. As only a small proportion of the sequenced individuals carry any given rare mutation, standard single variant association tests typically used to evaluate common variants in GWAS cannot be applied to rare variants because they will be severely underpowered unless the sample sizes or effect sizes are very large (Auer and Lettre 2015; Dering et al. 2012; Lee et al. 2014).

1.2 Rare variant analysis

To overcome the low power of rare variant studies, several statistical methods have been proposed to boost power, usually involving aggregating variants in a region and jointly testing the marginal effect of rare variants (Chen et al. 2011). These region-based tests can be broadly categorized into burden tests and non-burden tests. For burden tests, the cumulative effects of variants in a region are summarized into a single variable, which is then tested for association with the trait of interest. Variants can be collapsed by summing the number of risk alleles in a region or by using a dichotomous variable to indicate whether an individual carries any risk allele in the region of interest (Asimit et al. 2012; Li and Leal 2008; Madsen and Browning 2009; Morgenthaler and Thilly 2007; Morris and Zeggini 2010; Qi, Allen, and Li 2019). Burden tests are most powerful when variants in consideration are causal and the directions of the effect on the risk of the alternate/minor alleles are the same, either all positive or all negative. When the alternate/minor alleles of causal variants have effects on the phenotype in different directions, burden tests suffer from low power since the opposite effects of protective and deleterious alleles cancel out (Lee et al. 2014; Wang, Chen, and Yang 2012). To improve power in this situation, a data-adaptive sum test called aSum was proposed, where it allows for both trait-increasing and trait-decreasing variants in the model and a combined score is calculated from signs of the univariate tests (Han and Pan 2010). The downside of this method is that it is computationally expensive since p-values cannot be calculated analytically, thus requiring permutation.

To address the aforementioned issue of burden tests, non-burden tests have been proposed that focus on aggregating individual test statistics. Variance component tests evaluate the distribution of the variants within a region by aggregating the score statistics of the individual variants (Lin and Tang 2011; Lee et al. 2012; Svishcheva, Belonogova, and Axenovich 2014; Wu et al. 2011). Variance component tests are more powerful than burden tests when there are both positively and negatively associated variants in a region, but less powerful when the variants are in same direction (Basu and Pan 2011; Jiang and McPeck 2014; Lee et al. 2012). Among non-burden tests, the sequence kernel association test (SKAT) is a popular method that summarizes variant information using a kernel function and applies a variance component score test to evaluate the significance. SKAT is derived under the assumption that study participants are unrelated and treats the genotypes as random effects to reduce the number of parameters to estimate (Wu et al. 2011). Numerous extensions of SKAT have been proposed, such as famSKAT to allow for related individuals (Chen, Meigs, and Dupuis 2013), SKAT-O that combines burden and SKAT to maximize power (Lee et al. 2012), MONSTER that extends SKAT-O to account for familial correlation (Jiang and McPeck 2014), RC-SKAT that tests the cumulative effects of both common and rare variants (Ionita-Laza et al. 2013), and many more.

1.3 Gene-environment interaction

Both genetics and environmental factors play a role in etiology of complex diseases and thus, it has become our interest to also study GE interaction, as inclusion of GE may

provide insights into genetic and environmental influences on a trait of interest and provide better statistical models when both genetic and environmental influences correctly accounted for.

A general framework for a GE interaction model is the following:

$$g(E(y)) = B_0 + B_1G + B_2E + B_3EG,$$

where $g(\cdot)$ is a link function and y is the trait of interest. The null hypothesis for an interaction-only test is $H_0: B_3 = 0$ vs. the alternative hypothesis: $H_a: B_3 \neq 0$. We can also perform a joint test of the genetic variants and GE interaction: $H_0: B_2 = B_3 = 0$ vs. $H_a: B_2 \neq 0$ or $B_3 \neq 0$. A joint test evaluates whether the genetic variant is associated with the trait, allowing for interaction with an environmental variable. Joint tests are most powerful when both the genetic main effect and GE interaction effects are present. Some gene-based GE methods have been developed in the context of rare variants to reduce multiple testing burden (Chen, Meigs, and Dupuis 2014; Jiao et al. 2013; Tzeng et al. 2011).

1.4 Family data

In the early stages of GWAS, analyses focused on comparing cases with controls among unrelated individuals. Ignoring familial correlation and using linear regression to analyze family data leads to elevated type I error (Chen, Meigs, and Dupuis 2013). One way to resolve this issue is to select a subset of unrelated individuals, but this may substantially

reduce the sample size, thereby leading to power loss. In the case of rare variants, especially for GE interaction test, it is important to retain as many samples as possible to have appropriate power. A preferable method is to use a linear mixed model (LMM) to account for random polygenic effects that are shared within families and thus, eliminating the need to restrict to a subset of unrelated individuals.

1.5 Dichotomous and survival outcomes

Sometimes the trait of interest is not continuous. It could be binomially distributed, with the disease being present or absent, or time-to-event outcomes that are aiming to predict patients' risks for an adverse event during long-term follow up. The best-known method to analyze a dichotomous outcome is logistic regression, which is appropriate only when the observations are independent. To analyze dichotomous traits with family data in genetic studies, generalized linear mixed models (GLMMs) can be used to account for relatedness between samples. To analyze time-to-event outcomes, the Cox proportional hazards model is widely used (Cox 1972), and this model can be extended to include random effects using mixed effects Cox regression. Both GLMMs and mixed effects Cox regression have broad utility and are of great practical importance, but they have not been extensively used in the context of detecting GE interaction in rare variant analyses.

1.6 Dissertation outline

In this dissertation, we focus on statistical method development to detect GE interactions for rare variants for different types of outcomes. Each of the following three chapters

consist of methodological development, extensive simulation studies, and a real data application to the Framingham Heart Study (FHS). We then conclude by highlighting the contributions and future work.

In Chapter 2, we propose a rare variant by environment interaction test accounting for familial correlation under the GLMM framework (famGE). The framework can accommodate both binary and continuous traits with family data; so it is not necessary to restrict analyses to unrelated individuals.

In Chapter 3, we develop a GE interaction test for rare variants in studies of time-to-event outcomes, assuming the genetic main effects are fixed effects, while the GE interaction effects are random. We impose a ridge penalty on the genetic main effects to deal with potential multicollinearity between variants (coxGE). Given the importance of time-to-event outcomes in medical studies, we believe our proposed methods will be significant contributions in genetic association studies as well as time-to-event analyses.

In Chapter 4, we extend the coxGE framework to incorporate different types of penalties, such as least absolute shrinkage and selection operator (LASSO) and elastic net, and compare the performance of the model with coxGE with the ridge penalty proposed in Chapter 3 under various simulation settings.

Finally, in Chapter 5, we summarize the findings and outline future work.

Chapter 2 Methods for Detecting Gene by Environment Interaction in Rare Variant Analysis for Binary and Continuous Traits in Family Data

2.1 Introduction

Although GWAS have been successful in identifying genetic variants with strong association with disease and traits, variants evaluated in GWAS have been mostly restricted to common variants, typically defined as those with MAF greater than 1 or 5%. Additionally, these identified variants explain only a small portion of disease heritability, possibly be due to limited sample size and power in GWAS, and thus calls for performing meta-analysis (Riancho 2012). Nevertheless, even in large scale GWAS meta-analyses, much of the heritability remains unexplained. For example, in GWAS meta-analysis of adult height in > 93,000 East Asians, the investigators identified 98 loci at genome-wide significance that explain only about 9% of height heritability, which is a small proportion considering that human height heritability is approximately 80% (He et al. 2015).

One plausible explanation for unexplained heritability is the presence of low frequency and rare variants, which are not analyzed in GWAS due to their low MAF. It is well known that rare variants are responsible for many Mendelian disorders, but their roles have not been fully investigated in complex diseases (Ionita-Laza et al. 2013; Lee et al. 2014). With rapid advances and decreasing cost of whole-genome sequencing, attention has shifted to investigating the potential role of low frequency variants in complex human disease etiology. There is now an abundance of growing empirical evidence that rare

variants may be in part responsible for complex diseases, thus partially accounting for unexplained heritability (Kao et al. 2017; Lee et al. 2014; Yu et al. 2018) For example, Igartua et al. identified two novel rare variants associated with low-density lipoprotein cholesterol and high-density lipoprotein cholesterol with larger effects than the previously discovered variants within the known blood lipid associated loci (Igartua et al. 2017). In another study by He et al., multiple rare variants were found to be associated with lower systolic blood pressure (He et al. 2017). Successes from rare variant association studies highlight the importance of rare variants to complex disease susceptibility.

Another source of unexplained heritability could be due to GE interactions (Hamza et al. 2011; Matsui and Ehrenreich 2016). Complex diseases are multifactorial and involve both genetics and environmental factors. Therefore, only studying the main effects, either genetic or environmental, cannot provide full insights into the biological mechanisms and etiology of complex diseases. Methods for GE interactions for common variants have been well established (Chen, Meigs, and Dupuis 2014; Lin et al. 2013; Ma, Clark, and Keinan 2013; Moreno-Macias et al. 2010; Tzeng et al. 2011). To assess the interaction between a genetic variant and an environmental factor, we can test the interaction term itself, where we are solely interested in whether GE interaction is present, regardless of the significance of the genetic main effect, or jointly test both the main effect and the interaction term, in which case we are interested in determining if the genetic variant is associated with the phenotype, allowing for GE interaction. If GE interactions exist and

are correctly accounted for in the model, it will boost power to detect genetic signals. GE interaction methods developed for common variants will suffer from power loss if they are applied to rare variants. GE interaction analysis for rare variants is underdeveloped compared to common variants because it requires larger sample size to achieve comparable power.

Several GE interaction of rare variant methods are available but they are only applicable to unrelated individuals and cannot correctly account for familial correlation in their models (Lin et al. 2013; Su, Di, and Hsu 2017). Tzeng et al. developed the similarity-based regression method (SimReg) to test GE interaction effects of rare variants for continuous traits. It allows for covariate adjustments, models both main and interaction effects, and is computationally efficient. (Tzeng et al. 2011). Zhao et al. extended SimReg to allow for binary traits for both common and rare variants (Zhao et al. 2015). Lin et al. introduced a SNP-set GE interaction method using a variance component test under a generalized linear model framework (Lin et al. 2013). This method is developed for common variants but it can be easily extended to rare variants by applying weights to the variants. Chen et al. proposed two GE interaction tests (rareGE) and a joint test of main and interaction effects for rare variants using a variance components score test (Chen, Meigs, and Dupuis 2014). In rareGE, genetic variants can be included as fixed or random effects for the test of interaction and the method works for both binary and continuous traits. Mazo Lopera et al developed SNP-set GE interaction method for family data but it was proposed in the context of common variants (Mazo Lopera, Coombes, and

de Andrade 2017). Recently, Coombes et al. extended gene-based GE interaction methods to account for multiple interactions in family data but they are applicable to continuous outcomes only (Coombes 2018).

In this chapter, we develop a framework for testing GE interaction for rare variants called famGE to correctly incorporate family correlation. Our proposed approach can accommodate both binary and continuous traits in family data; so it is not necessary to restrict the analyses to unrelated individuals. We adopt a kernel-based method to leverage the joint information across the rare variants. We assume main effects of genetic variants, GE interaction term, and family correlation to be random effects in our model and implement a variance component score test in the GLMM framework to reduce the computational burden. When there are no related individuals, famGE will be equivalent to rareGE, a method proposed by Chen et al (Chen, Meigs, and Dupuis 2014). From our simulation studies, we show that famGE can control type I error, whereas rareGE has inflated type I error when familial correlation is not accounted for in family data.

This chapter is organized as follows. In section 2.2, we briefly introduce the GLMM framework, our notations, the GE interaction model (section 2.3), and the test statistic (section 2.4). In section 2.5, we conduct simulations under various settings to assess type I error rates and the power of our approach, comparing it to rareGE and the burden test. We illustrate an application of our method in testing gene by smoking interaction on body

mass index (BMI), using family data from the Framingham Heart Study (FHS) in section 2.6. We discuss our findings of our approach in section 2.7.

2.2 Generalized linear mixed model

Proposed by Breslow and Clayton, GLMMs are an extension of generalized linear model that can account for random effects (Breslow and Clayton 1993). They can accommodate a wide range of response distributions and a covariance matrix for the random effects.

Assuming a r -dimensional vector \mathbf{b} of random effects, y_i are conditionally independent with means $E(y_i|\mathbf{b}) = \mu_i$ and variances $var(y_i|\mathbf{b}) = \phi v(\mu_i)$, where ϕ is the dispersion parameter (1 for binary and Poisson data) and $v(\cdot)$ is the variance function. Denoting $\mathbf{y} = (y_1, \dots, y_n)^T$ and the design matrices with rows \mathbf{x}_i^T and \mathbf{z}_i^T by \mathbf{X} and \mathbf{Z} , the GLMM model is given by

$$\mathbf{g}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b}, \quad (1)$$

where \mathbf{b} is assumed to have a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{B} = \mathbf{B}(\boldsymbol{\sigma}^2)$ depending on an unknown vector $\boldsymbol{\sigma}^2$ of variance components.

2.3 Interaction test in generalized linear mixed model for family data

Assuming a sample of size n , let \mathbf{Y} be $n \times 1$ vector of phenotype (binary or continuous) with $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $var(\mathbf{Y}) = \phi v(\boldsymbol{\mu})$, where $\boldsymbol{\mu} = [\mu_1 \dots \mu_n]^T$ is the mean vector, ϕ is the dispersion parameter (1 for binary), and $v(\cdot)$ is the variance function. We consider the following GLMM for testing GE interactions:

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\alpha} + \mathbf{G}\mathbf{W}_1\boldsymbol{\theta} + \mathbf{E}\mathbf{G}\mathbf{W}_2\boldsymbol{\gamma} + \mathbf{d}, \quad (2)$$

where $g(\cdot)$ is the link function, \mathbf{X} is an $n \times p$ covariate matrix including the intercept and the environmental variable \mathbf{E} , $\boldsymbol{\alpha}$ is a $p \times 1$ vector associated with the fixed covariate effects, \mathbf{G} is an $n \times q$ genotype matrix, $\boldsymbol{\theta}$ is a $q \times 1$ vector of random effects for the genetic variants, $\mathbf{E}\mathbf{G}$ is an $n \times q$ GE interaction matrix, $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effects for GE interaction, and \mathbf{d} is an $n \times 1$ vector for the random effects of familial correlation. \mathbf{W}_1 and \mathbf{W}_2 are $q \times q$ diagonal matrices with pre-specified weights for genetic main effects and GE interaction effects, respectively. They measure genetic similarity between subjects via the genetic markers. Typically for rare variant analyses, we put higher weights to those variants that are rarer so the standard choice is to use a function of the inverse of the MAF of the genotypes as the weights. In famGE framework, user-defined weights can be flexibly included, such as weights calculated based on MAF or functional annotation scores. Good choices of weights can boost power (Wu et al. 2011). For the three random effects (genetic variants, GE interactions, and relatedness in families respectively), we assume that

$$\boldsymbol{\theta} \sim N(\mathbf{0}, \sigma_M^2 \mathbf{I}_q)$$

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_I^2 \mathbf{I}_q)$$

$$\mathbf{d} \sim N(\mathbf{0}, \sigma_G^2 \boldsymbol{\psi})$$

where $\boldsymbol{\psi}$ is twice the $n \times n$ kinship matrix from family relationships obtained from a pedigree or an empirical kinship calculated from genotype data to account for cryptic

relatedness. The kinship coefficient summarizes the genetic similarity between pairs of individuals. The random effects $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and \boldsymbol{d} are assumed to be independent. If $\boldsymbol{\gamma}$ is treated as a fixed effect, we would perform a q degrees of freedom score test, but this approach can suffer from power loss when q is moderate or large (Lin et al. 2013). By assuming $\boldsymbol{\gamma}$ follows a normal distribution with mean 0 and variance $\sigma_I^2 \mathbf{I}_q$, the null hypothesis for the interaction test: $H_0: \boldsymbol{\gamma} = \mathbf{0}$ is equivalent to testing $H_0: \sigma_I^2 = 0$ using a variance component score test (Lin et al. 2013). Score tests only require fitting the model under the null hypothesis, so they are more computationally efficient (Lin et al. 2016; Wu et al. 2011).

2.4 Estimation and hypothesis testing

To fit the null model for binary traits, we use the penalized quasi-likelihood method (Refer to Appendix A for complete derivation of the test statistic). This involves integrating over the random effects but this high dimensional integral is intractable, so we use Laplace approximation to estimate this integral. After the Laplace's method, we end up with the marginal likelihood that we can maximize with respect to our parameters.

We define the linear working vector under the null hypothesis $\mathbf{Y}_0 = \mathbf{X}\boldsymbol{\alpha} + \mathbf{G}\mathbf{W}_1\boldsymbol{\theta} + \boldsymbol{d} + \boldsymbol{\Delta}(\mathbf{Y} - \boldsymbol{\mu})$, where $\boldsymbol{\Delta} = \text{diag}\{g'(\mu_i)\}$, and let $\mathbf{D} = \text{diag}\left\{\frac{1}{\phi v(\mu_i)[g'(\mu_i)]^2}\right\}$. We iteratively fit the working vector to estimate the parameters until convergence to obtain our restricted maximum likelihood (REML) or maximum likelihood (ML) estimates (Breslow and Clayton 1993).

Let $\hat{\alpha}$, $\hat{\sigma}_M^2$, and $\hat{\sigma}_G^2$ be the REML (or ML) estimates under the null hypothesis. We can then calculate $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Delta}}$ and $\hat{\boldsymbol{D}}$ using the aforementioned REML or ML estimates. REML estimates the variance components independent of the fixed effects. ML produces unbiased estimation for the fixed effects but biased estimation of the variance components. In large samples, their results are usually close to each other. The restricted maximum quasi-likelihood function is defined as:

$$ql_R = -\frac{1}{2} \log |\boldsymbol{D}| - \frac{1}{2} \log |\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}| - \frac{1}{2} (\boldsymbol{Y}_0 - \boldsymbol{X}\boldsymbol{\alpha})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}_0 - \boldsymbol{X}\boldsymbol{\alpha}) \quad (3)$$

where $\boldsymbol{\Sigma} = \hat{\boldsymbol{D}}^{-1} + \sigma_M^2 (\boldsymbol{G}\boldsymbol{W}_1\boldsymbol{W}_1\boldsymbol{G}^T) + \sigma_G^2 \boldsymbol{\psi} + \sigma_I^2 (\boldsymbol{E}\boldsymbol{G}\boldsymbol{W}_2\boldsymbol{W}_2\boldsymbol{G}^T\boldsymbol{E})$.

To derive the score test for $H_0: \sigma_I^2 = 0$, we take the first derivative of equation 3 with respect to σ_I^2 ,

$$\frac{\partial ql_R}{\partial \sigma_I^2} = -\frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{E}\boldsymbol{G}\boldsymbol{W}_2\boldsymbol{W}_2\boldsymbol{G}^T\boldsymbol{E}) + \frac{1}{2} (\boldsymbol{Y}_0 - \boldsymbol{X}\boldsymbol{\alpha})^T \boldsymbol{\Sigma}^{-1} \boldsymbol{E}\boldsymbol{G}\boldsymbol{W}_2\boldsymbol{W}_2\boldsymbol{G}^T\boldsymbol{E}\boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}_0 - \boldsymbol{X}\boldsymbol{\alpha}) \quad (4)$$

The first term in equation 4 is fixed and independent of the phenotype. We follow the same rationale used in the derivation of the SKAT score statistic and take twice the second term to be our test statistic (Chen, Meigs, and Dupuis 2013; Wu et al. 2011)

$$\boldsymbol{Q} = (\hat{\boldsymbol{Y}}_0 - \boldsymbol{X}\hat{\boldsymbol{\alpha}})^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{E}\boldsymbol{G}\boldsymbol{W}_2\boldsymbol{W}_2\boldsymbol{G}^T\boldsymbol{E}\hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{Y}}_0 - \boldsymbol{X}\hat{\boldsymbol{\alpha}}) \quad (5)$$

where $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{D}}^{-1} + \hat{\sigma}_M^2 (\boldsymbol{G}\boldsymbol{W}_1\boldsymbol{W}_1\boldsymbol{G}^T) + \hat{\sigma}_G^2 \boldsymbol{\psi}$.

Under the null hypothesis $H_0: \sigma_l^2 = 0$, $\mathbf{Q} \sim \sum_{j=1}^q \lambda_j \chi_{1,j}^2$, where λ_j 's are eigenvalues from the matrix $\mathbf{W}_2 \mathbf{G}^T \mathbf{E} (\widehat{\Sigma}^{-1} - \widehat{\Sigma}^{-1} \mathbf{X} (\mathbf{X}^T \widehat{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\Sigma}^{-1}) \mathbf{E} \mathbf{G} \mathbf{W}_2$ (Wu et al. 2011; Zhang and Lin 2003).

2.5 Simulation Studies

2.5.1 Type I Error

2.5.1.1 Type I Error Simulation Settings

We performed simulation studies to evaluate the performance of our method for both continuous and binary phenotypes. For the null simulation study, we first considered the scenario where there is genotype main effect but no GE interaction effect in simulating our phenotypes. To simulate the genotypes, we used the SeqSIMLA software, which can simulate sequence data in families with user-specified pedigree structures. We used the reference sequence based on 1000 Genomes Project for European populations (Chung and Shih 2013). We picked 7 families from FHS with family membership ranging from 120 to 640 (2030 individuals in total). In the simulated genotypes, we chose a region that spans from 1,100 base pairs to 1,140 base pairs on chromosome 1. To simulate our phenotype, we varied the proportion of low frequency (MAF < 5%) causal SNPs included in our model from 20% to 40% to 60% and 80% for each of 20,000 replicates. We considered both continuous and binary phenotypes. For each of the 20,000 replicates, we simulated phenotype datasets from

$$\mathbf{y} = 0.1 + 0.1\mathbf{age} + 0.5\mathbf{sex} + 0.3\mathbf{smoke} + \mathbf{d} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where *age* was generated from a normal distribution with mean of 50 and standard deviation equal to 5, *sex* was generated from a Bernoulli distribution with probability 0.5, *smoke* was generated from a Bernoulli distribution with probability 0.5, and ϵ was generated from a standard normal distribution. Family correlation, \mathbf{d} , is from a multivariate normal distribution with means 0 and covariance $\sigma_G^2 \boldsymbol{\psi}$, where σ_G^2 is set to 1 and $\boldsymbol{\psi}$ is twice the kinship matrix. $\boldsymbol{\gamma}$ consist of effect sizes for the causal SNPs. For binary traits, we first simulated the continuous traits and set the lower 80% as controls (0's) and the upper 20% as cases (1's). We simulated variants where the directions of the genetic main effect (represented by $\boldsymbol{\gamma}$) on the risk of the minor allele are either all positive or mixed with 50% positive and 50% negative and the effect sizes were determined by

$$\gamma_i = \frac{h}{2MAF_i(1 - MAF_i)}$$

where MAF is the minor allele frequency of SNP i and h is a constant calculated as

$$h = \frac{R^2}{\mathbf{v}^T \mathbf{L} \mathbf{v}}$$

and R^2 , the proportion of variance explained by the causal SNPs, is fixed at 1% for causal SNPs with effect sizes in same direction and 5% for causal SNPs with effect sizes in opposite directions. The correlations between the SNPs are in matrix \mathbf{L} , and \mathbf{v} is a vector that indicates the direction of the SNP effects.

We performed three tests for the type I error comparison: our proposed method that correctly accounts for familial correlation (famGE), rareGE where the main genetic

effects are treated as random and family correlation is ignored (rareGE), and the burden test using GLMMs to account for familial correlation (BT). For the burden test, we used an indicator of whether or not at least one rare allele was present in the testing region. We simulated 20,000 replicates and used Wu weights, which are the beta density function with parameters 1 and 25 evaluated at the MAF of the variants, for famGE and rareGE tests (Wu et al. 2011).

2.5.1.2. Simulation Results for Type I error

Table 1 includes the results for type 1 error results for the test of interaction only for famGE, rareGE, and BT at significance levels, α of 0.05, 0.01, and 0.001 from 20,000 simulation replicates. Both famGE and BT have correct type 1 error rates at all three α levels for both continuous and binary traits. When familial correlation is not appropriately taken into account in the model, rareGE test suffers from type 1 error inflation, which is more pronounced in binary traits. famGE has valid type 1 error rates under various scenarios, such as differing the direction of main genotype effects and increasing the proportion of causal variants in the model for both continuous and binary traits, although it is slightly conservative. This is similar to the result observed using SKAT to test the main genetic effects for dichotomous traits (Wu et al. 2011). Because rareGE does not have correct type 1 error rate when familial correlation is not taken into account, we did not include rareGE in the power comparison in the next section.

Table 1: Comparison of famGE, rareGE, and BT type 1 errors based on 20,000 replications (in %)

Scenarios (+/-/0)	α	Continuous			Binary		
		famGE ^a	rareGE ^b	BT ^c	famGE ^a	rareGE ^b	BT ^c
4/0/16	5	4.96	5.30	5.01	5.03	5.51	5.07
	1	1.00	1.10	1.02	1.02	1.25	1.05
	0.1	0.12	0.17	0.10	0.12	0.18	0.13
8/0/12	5	4.88	5.21	4.99	4.84	5.57	4.92
	1	1.02	1.08	1.03	0.97	1.34	0.95
	0.1	0.11	0.17	0.12	0.11	0.19	0.09
12/0/8	5	4.95	5.20	4.99	4.82	5.54	4.98
	1	0.94	1.20	0.96	1.06	1.22	0.90
	0.1	0.12	0.15	0.11	0.08	0.15	0.105
16/0/4	5	5.04	5.38	5.00	4.84	5.70	5.00
	1	1.01	1.16	0.99	1.04	1.41	0.94
	0.1	0.11	0.16	0.098	0.12	0.20	0.10
2/2/16	5	4.80	5.25	5.00	4.97	5.56	5.06
	1	0.95	1.20	1.01	0.94	1.16	0.97
	0.1	0.10	0.15	0.11	0.11	0.16	0.11
4/4/12	5	4.99	5.33	5.02	4.84	5.52	4.89
	1	0.93	1.05	1.00	0.97	1.16	0.96
	0.1	0.099	0.19	0.10	0.10	0.14	0.09
6/6/8	5	4.93	5.41	4.94	4.88	5.61	4.88
	1	1.03	1.10	1.04	0.92	1.37	1.00
	0.1	0.097	0.14	0.103	0.104	0.17	0.11
8/8/4	5	5.02	5.31	4.93	4.89	5.55	4.79
	1	0.90	1.11	0.99	0.97	1.27	0.94
	0.1	0.10	0.17	0.12	0.105	0.16	0.10

+/-/0: number of variants whose main genotype effects are positive/ negative/ neutral

α is expressed in percentage

famGE^a: our proposed method accounting for family correlation

rareGE^b: interaction test ignoring family correlation

BT^c: burden test

2.5.2 Power

2.5.2.1 Power Simulation Settings

To assess power, we simulated data under the alternative hypothesis, where we include a gene by smoking interaction effect, in addition to the genotype main effects. Similar to the type I error simulation, genotypes were simulated using the SeqSIMLA software and we selected SNPs with MAF less than 5% and varied the directions of genetic main effect and the proportions of causal SNPs included in the model. We simulated 10,000 phenotype datasets from

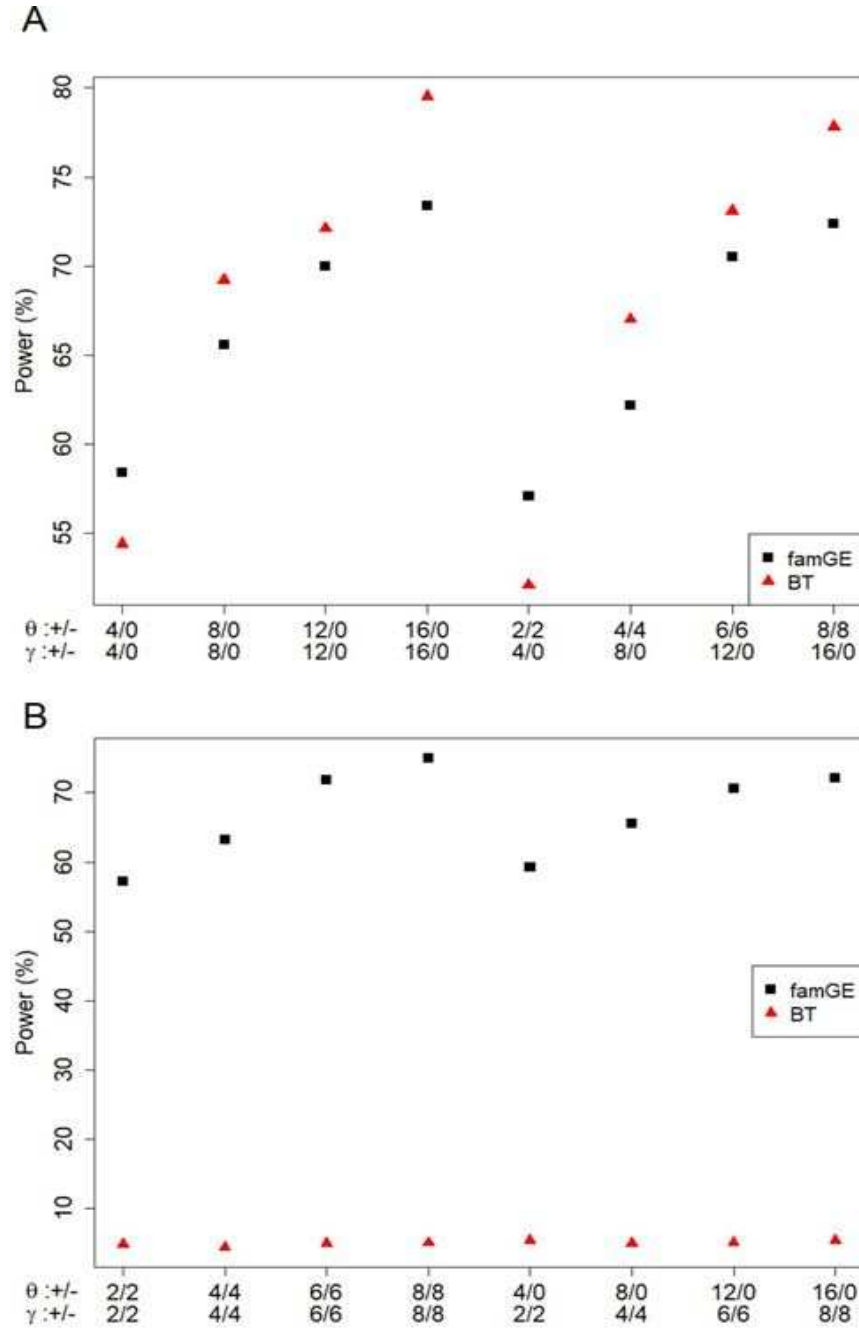
$$y = 0.1 + 0.1age + 0.5sex + 0.3smoke + d + G\gamma + (smoke - 0.5)G\delta + \epsilon,$$

where age, sex, and smoke, family correlation, and error terms were generated from the same distribution described in the type I error simulation study and the genotype effects γ were determined the same way as in our null simulation study. Interaction effects δ were generated from a normal distribution with mean 2 and standard deviation of 0.3. We considered the scenarios where the directions of the interaction of a causal variant (represented by δ) is either the same or opposite to the directions of the corresponding main effect (represented by γ). Negative interaction effects were simulated the same way as above except we multiplied the effects by -1. To test binary outcomes, we set the lower 80% of the simulated continuous outcome to be the controls and the upper 20% to be the cases. For power comparison, we also performed a burden test, where the summary variable for each individual was created using an indicator of whether or not at least one rare allele was present in the testing region.

2.5.2.2 Simulation Results for Power of the test of interaction only

Figure 1 shows the power comparison for famGE and BT for continuous traits at $\alpha = 0.001$ from 10,000 replicates. When the proportion of causal variants is low, BT has lower power compared with famGE (1A). This is expected because burden tests are powerful when a large proportion of causal variants are included in the region. When the proportion of causal variants increases, we see that powers for both BT and famGE increase but we see a larger power increase in BT compared to famGE. However, when variants have interaction effects in different directions, we observe a power drop, with power close to 0 for BT, whereas famGE shows increasing power as the proportion of causal variants increases (1B). FamGE maintains fairly consistent power across different scenarios, regardless of the direction of the main effects.

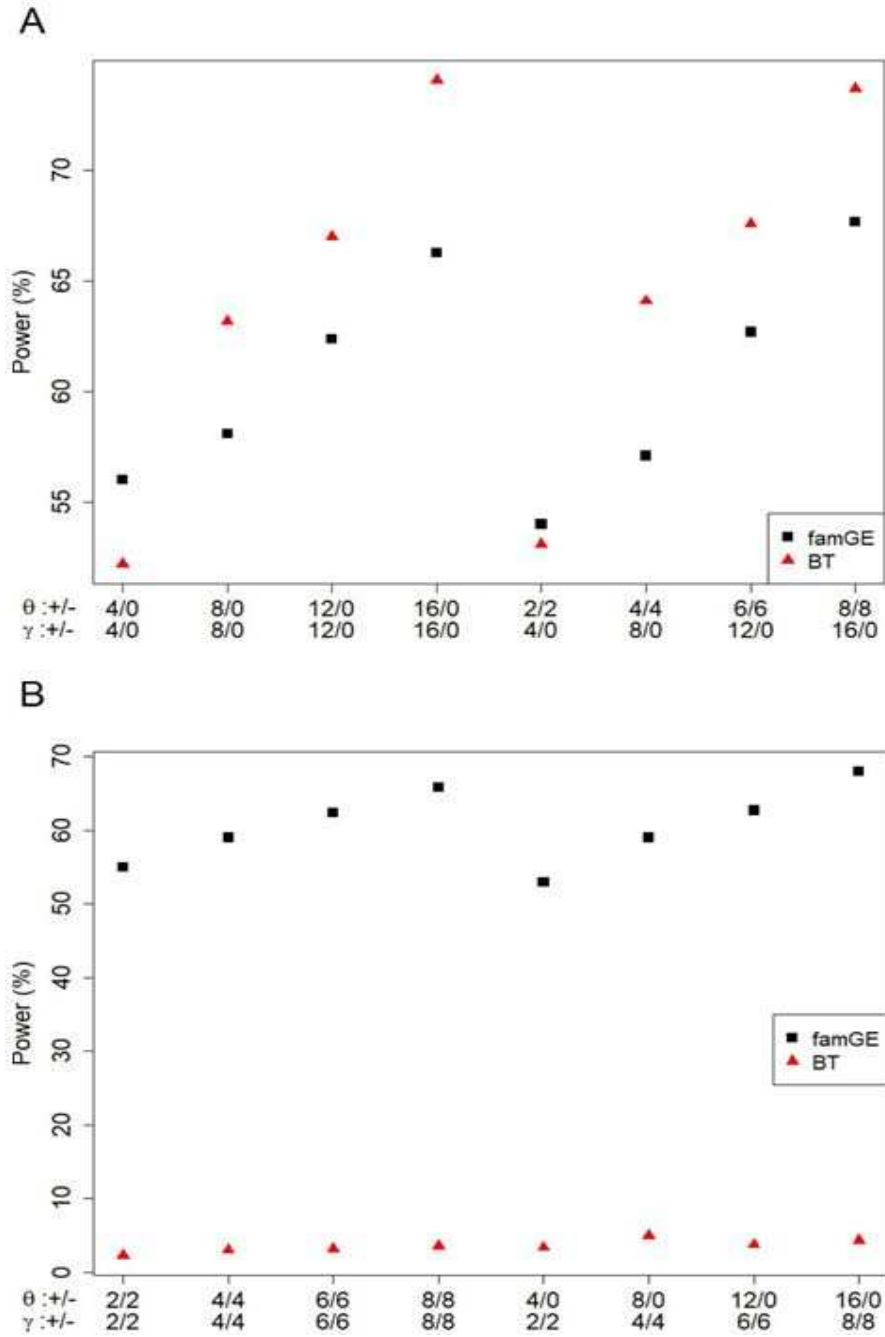
Figure 2 shows power comparisons of famGE and BT for binary traits at $\alpha = 0.001$ from 10,000 replicates. We come to the same conclusion as the results seen from continuous traits. Even though burden tests have higher power than famGE in the case where interaction effects are in the same direction, with the exception where there are 4 causal variants and 16 neutral variants in the model, famGE is able to maintain high power regardless of the direction or the proportion of variants (2A). The burden test significantly loses power in the presence of both trait-increasing and decreasing alleles, whereas performance of famGE is robust across different scenarios (2B).

Figure 1: Power comparison of famGE and BT for continuous trait

(A) All positive interaction effects

(B) Minor alleles with both positive and negative interaction effects

+/- indicates the number of SNPs with positive and negative effects. θ denotes SNP main effects and γ denotes GE interaction effects

Figure 2: Power comparison of famGE and BT for binary trait

2.6 Application to the Framingham Heart Study

In the real data application, we illustrate our method to test gene-based interaction with smoking on a quantitative trait, BMI, and a dichotomous trait, overweight status ($BMI \geq 25$), using participants from the FHS.

Obesity is a world-wide problem that can lead to serious health problems, such as high blood pressure, type 2 diabetes, heart disease, and more. Conventionally, BMI has been used as a way of measuring obesity. GWAS have identified numerous loci that are associated with BMI, but whether these genetic effects are modulated by environmental factors have not been extensively investigated (Liao et al. 2016; Rask-Andersen et al. 2017). Recently, Justice et al. investigated the effect of smoking on genetic susceptibility to obesity in a large consortium meta-analysis of 241,258 individuals and reported two common variants reaching genome-wide significant threshold, rs12902602 and rs336396 near *CHRNBA* and *INPP4B* genes respectively (Justice et al. 2017). Here we evaluate whether there are modification effects of smoking status on genetic risk from these two loci for obesity with rare or less frequent variants.

We analyze genotype data from the Illumina V1.0 Exome Chip and select variants with MAF less than 5%. We adjust for age, sex, cohort (four category variable), first two principal components, and smoking main effect in our model and pedigree-based kinship matrix to account for familial correlation. Our data consists of 596 individuals from the Original Cohort (Exam 21), 2547 from the Offspring Cohort (Exam 8), 3868 from the Gen

3 Cohort (Exam 1), and 177 from the Omni Cohort (Exam 1). There are 3264 males (45.4% males), 6063 non-smokers (84.4% non-smokers), and their ages range from 19 to 85 (median = 49). We test for gene x smoking interaction by treating BMI as a continuous trait or as a binary trait by dichotomizing BMI at 25, which classified 4502 individuals as overweight and 2686 individuals in the normal range. A total of 7188 individuals are included in the gene x smoking interaction analysis.

We consider two genes, Cholinergic Receptor Nicotinic Beta 4 (*CHRNB4*) and Inositol Polyphosphate-4-Phosphatase (*INPP4B*). Table 2 summarizes the analysis results for these two genes using our proposed method and the burden test. In order to minimize the inclusion of potential noise in our testing region, we only include variants with MAF < 5% that are annotated as either stop-gain/loss, splice, or missense. We can achieve higher statistical power by utilizing functional annotation to prioritize variants predicted to have potential biological significance. With the burden test, we find a significant *CHRNB4* x smoking interaction for overweight status (p-value = 0.0184) at $\alpha = 0.025$. Using famGE, we find *CHRNB4* to be statistically significant at $\alpha = 0.025$ for both continuous (p-value = 0.0063) and dichotomized (p-value = 0.0023) BMI, but no gene x smoking interaction was identified for gene *INPP4B*. (See Appendix C for quantile-quantile plot of famGE applied genome-wide.) For *CHRNB4*, we notice that the signal is stronger in the binary model, which could potentially indicate a non-linear interaction between smoking and *CHRNB4*.

From the two genes we test in the real data application, *CHRNA4* shows statistical significance in interacting with smoking on BMI. The *CHRNA4* gene has been reported to be associated with higher BMI in never smokers and lower BMI in current smokers, implying that genetic variants may influence BMI via the weight-reducing effects of smoking in opposite directions (Taylor et al. 2014). The *INPP4B* gene is a novel locus identified in the meta-analysis by Justice et al. (Justice et al. 2017) but to our knowledge, this finding has not yet been replicated in other studies. We cannot exclude the possibility that there are no rare variants interacting with smoking in *INPP4B* gene (Justice et al. 2017). However, it is also possible that reduced power due to limited sample size in our study compared to that in the meta-analysis restrict us from finding a significant association.

Table 2: Association results of gene by smoking interaction on continuous trait: BMI and binary trait: overweight status ($BMI \geq 25$)

Gene	Chromosome	# of variants included	famGE p-value	Burden test p-value
Continuous: BMI				
<i>CHRNA4</i>	15	8	0.0063	0.3787
<i>INPP4B</i>	4	5	0.5504	0.7797
Binary: $BMI \geq 25$				
<i>CHRNA4</i>	15	8	0.0023	0.0184
<i>INPP4B</i>	4	5	0.9166	0.9952

of variants included refer to functionally relevant variants (stop-gain/loss, splice, or missense) that were available on the exome array within each of the genes

2.7 Discussion

In this chapter, we develop a method called famGE in detecting GE interactions of a set of rare variants using GLMM. This proposed approach can accommodate both binary and continuous traits in family data or samples with cryptic relatedness. Additionally, famGE allows weighting variants differently based on prior information such as allele frequency. Under this model, we treat the genetic variants, familial correlation, and GE interaction effects as random effects and implement a variance component score test, which only requires fitting the null model, and thus, reduces computational burden. Our simulation studies show that famGE maintains correct type 1 error and high power under various scenarios. Another attractive feature of famGE is that it can calculate p-values without the need for permutation.

Ignoring familial correlation when using linear or logistic regression to analyze family data lead to inflated type I error (Chen, Meigs, and Dupuis 2013). One way to resolve this issue is to select a subset of unrelated individuals, but this may substantially reduce the sample size and lead to power loss. In the case of rare variants, especially for GE interaction test, it is important to retain as many samples as possible to have appropriate power. A preferable method is to use GLMMs that account for familial correlation as a random effect and thus, eliminates the need to restrict to unrelated individuals. In famGE, kinship coefficients can be obtained either from a pedigree or an empirical kinship matrix. Because the empirical kinship matrix can account for cryptic relatedness, it is

advantageous to use the empirical kinship to estimate the level of relatedness among individuals (Wu et al. 2011)

The proposed famGE method models the genotype main effects as random effects in order to reduce the number of parameters that need to be estimated. If one wishes to model the genotypes main effects as fixed effects, the derivations of the new test statistic follows the same framework as the test statistic for famGE with genotype main effects as random (Refer to Appendix B for the derivation of famGE with genotype main effects as fixed). When the number of variants included in the model is large, however, there is a potential for multicollinearity with the covariates and/or among the variants when fitting the model. It is also shown that modeling the genetic main effects as fixed leads to slightly inflated type 1 error rate at less stringent α levels when the number of variants included in the testing region is large (Chen, Meigs, and Dupuis 2014). Therefore, modelling genetic main effects as random effects is preferable.

Advanced development and decreasing cost of sequencing technology have facilitated the discovery and accessibility to low frequency variants and there is growing evidence that they are implicated in complex diseases. Therefore, more attention has been brought to analyzing and developing rare variants methods. GE interaction may explain part of the unexplained heritability, provide insight into etiology of disease, identify subgroups in the population that are at high risk, and help develop personalized treatments. Our

proposed approach is flexible in that it can accommodate either binary or continuous traits in related samples.

Chapter 3 Methods for Detecting Gene by Environment Interaction in Rare Variant Analysis for Time-to-Event Outcomes

3.1 Introduction

Many medical investigations are interested in predicting patients' risks for an adverse outcome during long-term follow up, and time-to-event data are informative because they contain more information than whether an event occurred or not. As a result, there has been an increased interest to identify genetic markers that are predictive of patient's prognosis. Such markers can help identify those patients who may benefit from receiving earlier and/or more aggressive treatments to hopefully improve their survival.

Additionally, discovery of these prognostic markers can help elucidate the underlying biological mechanisms involved in disease progression and eventually lead to developing personalized treatments. To identify markers associated with a patient's prognosis, investigators will usually employ a prospective cohort design to follow subjects and collect covariates from baseline to the time when the event, such as death or recurrence of disease, occurs.

Time-to-event data are unique because the outcome is not just whether or not an event occurred but also how much time is between the start of follow-up and the event/censoring (survival time). In survival analysis, linear regression is not suitable to model the survival times as a function of covariates for two main reasons. First, linear regression is not equipped to handle censoring of observations, in which their exact

survival time is incomplete. Second, survival times most often have a skewed distribution, so linear regression is not appropriate unless the survival times are transformed such that they are normally distributed.

To date, numerous methods for assessing rare variant association exist for binary and continuous outcomes, but limited work has been done for analyzing time-to-event outcomes for rare variants (Chen et al. 2015; Leclerc et al. 2015; Lin et al. 2011). Cai et al. developed a kernel machine method to test for pathway effects based on gene expression data on survival outcomes (Cai, Tonini, and Lin 2011). Lin et al. extended this method to test a set of common genetic variants on time-to-event outcomes (Lin et al. 2011). Chen et al. proposed a likelihood ratio test statistic to analyze survival outcomes using the Cox proportional hazard model (Chen et al. 2015). These three methods were developed for unrelated individuals and cannot appropriately model familial correlation. Leclerc et al. proposed a SNP-set association test for censored traits, with adjustment for familial relatedness (Leclerc et al. 2015). Recently, Qi et al. proposed a family-based burden and kernel test for analyzing censored traits (Qi, Allen, and Li 2019). All of these aforementioned methods were developed to test the genotype main effects only without accounting for GE interactions. To the best of our knowledge, there are no existing methods available in the literature to date for testing the association between GE interactions for a set of rare variants and time-to-event outcomes.

To address this gap in the literature, in this chapter, we propose a test for the association between GE interaction of rare variants and censored time-to-event outcomes (coxGE). We use the kernel machine Cox regression framework to assess the joint effects of the rare variants on the survival outcome and implement a variance component score test within the Cox proportional hazards model. Under this model, the genotype main effects are treated as fixed effects, while the GE interactions are modeled as random effects. As some variants may be highly correlated due to high linkage disequilibrium (LD), we impose a ridge penalty to shrink the size of the coefficients for the genetic main effects in the null model.

This chapter is organized as follows. In sections 3.2 - 3.4, we briefly introduce survival analysis, the Cox proportional hazards model, penalized regressions and the ridge penalty that will be used in our model. In section 3.5, we describe our GE interaction model and the test statistic. In section 3.6, we conduct extensive simulations under various settings to assess type 1 error rates and power of our approach. We illustrate our approach in testing gene by smoking interaction on time-to-fracture with samples from the Framingham Osteoporosis Study in section 3.7. We summarize our main findings and discuss future work in section 3.8.

3.2 Survival analysis

Survival analysis, also known as time-to-event analysis, refers to the analysis of data where the outcome is the duration of time until the occurrence of an event, such as death,

relapse, or disease occurrence. Survival analysis is not only used in biomedical sciences, but in various fields such as engineering, finance, and sociology to evaluate time until equipment failure, time until market crash, and event history analysis, etc. The goal of survival analysis is to understand the underlying distribution of time elapsed from a starting point to the occurrence of an event, known as survival time, and to assess the relationship between survival time and the independent variables (covariates) (Clark et al. 2003; Stel et al. 2011). By the end of the study, not all subjects will have experienced the event, and therefore survival times for that subset of the study group are unknown. For these individuals, where their time-to-event information is unknown and incomplete, they are said to be censored, which is a unique characteristic that distinguishes survival analysis from other areas in statistics.

As mentioned before, one challenge for survival analysis is that only a subset of the individuals will experience the event, while the rest are censored. The most common form of censoring is right-censoring, which happens due to either the participants dropping out of the study and therefore lost to follow up or they have not yet experienced the event before the end of study observation time. When an individual is censored, we do not know the true survival time for that participant. Therefore, censoring represents a type of missing data that is assumed to be random and non-informative, which implies that censoring is unrelated to the probability of an event occurring. Simply excluding censored observations from the data could bias the results. Unlike linear or logistic regression, Cox proportional hazards model can incorporate both the censored and un-

censored individuals in the model to estimate the parameters of interest (George, Seals, and Aban 2014; Clark et al. 2003; Stel et al. 2011).

3.3 Cox proportional hazards model

A model that is commonly used to evaluate the association between several clinical/risk factors and survival time of independent study participants is the Cox proportional hazards model (Cox 1972). The Cox model assumes a semi-parametric form for the hazard, and the parameter estimates can be obtained by maximizing a partial likelihood. The Cox model assumes that the covariates have a multiplicative effect on the hazard and can be written as the following:

$$h(t) = h_0(t)e^{\mathbf{X}\boldsymbol{\beta}},$$

where t represents time, $h(t)$ is the hazard function determined by a set of covariates \mathbf{X} , $\boldsymbol{\beta}$ are the effect sizes of the covariates, and $h_0(t)$ is the baseline hazard if all the covariates \mathbf{X} are 0. We make a parametric assumption for the effect of \mathbf{X} on the hazard functions, while no assumption is made for $h_0(t)$, thus resulting in a semi-parametric property for the Cox proportional hazards model. The only requirement for the baseline hazard is that $h_0(t) > 0$ (Chen, Ibrahim, and Shao 2018; Flynn 2012; George, Seals, and Aban 2014; Stel et al. 2011).

Inference can be made via the partial likelihood instead of the full likelihood to remove the nuisance parameter $h_0(t)$

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \left\{ \frac{e^{X_i^T \boldsymbol{\beta}}}{\sum_{j \in R_i} e^{X_j^T \boldsymbol{\beta}}} \right\}^{\Delta_i}$$

where R_i contains those who are at risk at time t_i (i.e. people who are still alive up to time t_i) and Δ_i is the event indicator (1 for those who experience the event and 0 for those who are censored). Since the baseline hazard function is left completely unspecified, standard maximum likelihood methods cannot be used to estimate $\boldsymbol{\beta}$. Using the Cox partial likelihood, we are able to estimate $\boldsymbol{\beta}$ while ignoring the baseline hazard function completely. It has been shown that the parameters estimated from the partial likelihood have the same properties as ones derived from the full likelihood (Chen, Ibrahim, and Shao 2018; Simon et al. 2011).

The Cox partial likelihood defined above is determined by the order in which events occur and not the actual times at which they occur. Therefore, it assumes that there are no tied events among the observations, and thus the event times can be uniquely sorted. Even though time is measured on a continuous scale, event times are usually reported to the nearest day or week, so ties can be present in many data sets, which is problematic. For example, if two subjects A and B have the same event time, it is unclear whether subject A should be in the risk set while subject B is experiencing the event and vice versa. Since the Cox regression model uses the ranks of event times to fit the model, tied events create problems and the above partial likelihood function will need to be adjusted in order to take into account of tied events. Several modified partial likelihood functions have been

proposed, such as the Exact method (Allison 2011; DeLong, Guirguis, So 1994), the Breslow method (Breslow 1974), the Efron approximation (Efron 1977), and the Discrete method (Cox 1972) for handling tied event times.

3.4 Regularized regression

In the classic case where the sample size is greater than the number of covariates, the Cox model performs well. When the number of covariates, such as genetic variants, exceeds the sample size, however, it leads to degenerate behavior. To address the high-dimensionality and potential multicollinearity issues among the predictors and avoid overfitting, one can use regularized regression to add a constraint/penalty to the residual sum of squares. The consequence of imposing this penalty is to shrink the coefficient estimates towards 0 to discourage fitting a complex model. This penalty term, referred to as shrinkage, has a tuning parameter λ that controls the amount of penalization. As λ increases, the coefficient estimates shrink towards 0. Most often, λ is determined by K-fold cross validation, and usually the value of λ that gives the minimum cross-validated error is selected.

Many different types of penalization methods have been proposed to handle the case with $p > n$. In our model, we will use the ridge penalty to prevent the coefficients from becoming too large. Ridge regression adds a penalty term on the squared L_2 norm of the coefficient vector to shrink the coefficients towards 0. Ridge regression is

computationally efficient and a good default regularization method in the presence of correlated variables (Hoerl and Kennard 1970).

3.5 Gene-environment interaction test for survival outcomes

Let $\mathbf{X} = [\mathbf{X}_1 \dots \mathbf{X}_n]^T$ be a $n \times p$ matrix of covariates including the environmental variable of interest \mathbf{E} , $\mathbf{G} = [\mathbf{G}_1 \dots \mathbf{G}_n]^T$ be a $n \times q$ genotype matrix, and $\mathbf{EG} = [\mathbf{EG}_1 \dots \mathbf{EG}_n]^T$ be an $n \times q$ GE interaction matrix. Let T denote the survival time and C be the censoring time. Due to censoring, T is observable up to a bivariate vector (U, Δ) , where $U = \min(T, C)$ is the observed time and $\Delta = I(T \leq C)$ is the event indicator (1 for those who experience the event and 0 for those who are censored). We assume C is independent of T , conditional on $\mathbf{X}, \mathbf{G}, \mathbf{EG}$. We can relate T to $\mathbf{X}, \mathbf{G}, \mathbf{EG}$ using Cox proportional hazards model defined as:

$$h(t) = h_0(t)e^{\mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{W}_1\boldsymbol{\theta} + \mathbf{EG}\mathbf{W}_2\boldsymbol{\gamma}},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector associated with the fixed covariate effects, $\boldsymbol{\theta}$ is a $q \times 1$ vector associated with fixed genetic main effects, $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effects for GE interaction and \mathbf{W}_1 and \mathbf{W}_2 are $q \times q$ diagonal matrices with pre-specified weights for genetic main effects and GE interaction effects respectively. In the coxGE framework, user-defined weights can be flexibly included, such as weights calculated based on MAF or functional annotation scores. For the GE random effect, we assume that

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_G^2 \mathbf{I}_q)$$

We are interested in testing GE interaction of rare variants. This corresponds to testing $H_0: \boldsymbol{\gamma} = \mathbf{0}$ which is equivalent to testing $H_0: \sigma_G^2 = 0$ using a variance component score test (Lin et al. 2013; Wu et al. 2011). Score tests only require fitting the model under the null hypothesis, so they are more computationally efficient. We can write the Cox partial likelihood of the hazard function as the following:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^n \left\{ \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{G}_i^T \mathbf{W}_1 \boldsymbol{\theta} + E_i \mathbf{G}_i^T \mathbf{W}_2 \boldsymbol{\gamma}}}{\sum_{j \in R_i} e^{\mathbf{X}_j^T \boldsymbol{\beta} + \mathbf{G}_j^T \mathbf{W}_1 \boldsymbol{\theta} + E_j \mathbf{G}_j^T \mathbf{W}_2 \boldsymbol{\gamma}}} \right\}^{\Delta_i}$$

where R_i contains the set of independent individuals who are at risk at time t_i (people who are still alive and free of event up to time t_i)

The log partial likelihood with respect to $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}$ is:

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^n \Delta_i \left\{ \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{G}_i^T \mathbf{W}_1 \boldsymbol{\theta} + E_i \mathbf{G}_i^T \mathbf{W}_2 \boldsymbol{\gamma} - \log \left(\sum_{j \in R_i} e^{\mathbf{X}_j^T \boldsymbol{\beta} + \mathbf{G}_j^T \mathbf{W}_1 \boldsymbol{\theta} + E_j \mathbf{G}_j^T \mathbf{W}_2 \boldsymbol{\gamma}} \right) \right\}$$

When the number of genetic variants, \mathbf{G} , included in the model is large, there is potential for multicollinearity among the variants due to possible high LD, which will lead to unstable parameter estimates and difficulty in interpretation. Therefore, we impose a penalty on $\boldsymbol{\theta}$ to shrink the estimated coefficients towards 0. We implement ridge regression to add a L2-norm penalty term to the loss function, which is the squared magnitude of the coefficient. This regularization forces the variables with minor

contribution to the model to be close to 0 but not exactly equal to 0, no matter how big we set the λ to be.

We can write the log partial likelihood with the ridge penalty as the following:

$$l_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) - \lambda \sum_{j=1}^q \theta_j^2$$

We can see that λ controls the strength of shrinkage. If $\lambda = 0$, we are left with regular Cox partial likelihood and no penalty will be imposed on the variables and as $\lambda \rightarrow \infty$, all coefficient estimates will be very close to 0 but never actually equal to 0 since ridge regression does not perform variable selection.

Similar to the calculations in Chen et al (2015), we can write our partial derivatives as:

$$\frac{\partial l_p}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \mathbf{M}$$

$$\frac{\partial l_p}{\partial \boldsymbol{\theta}} = \mathbf{W} \mathbf{G}^T \mathbf{M} + 2\lambda(\boldsymbol{\theta}_j)$$

$$\frac{\partial l_p}{\partial \boldsymbol{\gamma}} = \mathbf{W} \mathbf{G}^T \mathbf{E} \mathbf{M}$$

where we define \widehat{M}_i to be Martingale's residual under the null as the following:

$$\widehat{M}_i = \Delta_i - \int_0^\infty I(U_i \geq t) e^{\mathbf{X}_i \widehat{\boldsymbol{\beta}} + \mathbf{G}_i \mathbf{W}_1 \widehat{\boldsymbol{\theta}}} d\widehat{\Lambda}_0(t) = \Delta_i - \epsilon_i$$

Here, $\widehat{\Lambda}_0(u) = \frac{\sum_{i=1}^n \Delta_i I(U_i \leq u)}{\sum_{i=1}^n \Delta_i I(U_i \geq t) e^{X_i \beta + G_i W_1 \theta}}$ is Breslow's estimator of $\Lambda_0(u)$ (baseline hazard function) under the null hypothesis of no GE interaction.

Then, we integrate out the random effect to obtain the log integrated partial likelihood with respect to β, θ , and σ_G^2 :

$$\begin{aligned} \widetilde{l}_p &= \log \int e^{l_p(\beta, \theta, \gamma)} dF(\gamma; \sigma_G^2) \\ \widetilde{l}_p &= \log \int \left(e^{l_p(\gamma=0)} + e^{l_p(\gamma=0)} \frac{dl_p(\gamma=0)}{d\gamma^T} \gamma + \frac{1}{2} e^{l_p(\gamma=0)} \gamma^T \left(\frac{dl_p(\gamma=0)}{d\gamma} \frac{dl_p(\gamma=0)}{d\gamma^T} + \frac{d^2 l_p(\gamma=0)}{d\gamma d\gamma^T} \right) \gamma \right) dF(\gamma; \sigma_G^2) \\ &= l_p(\gamma=0) + \log \left(1 + \int \frac{1}{2} \gamma^T \left(\frac{dl_p(\gamma=0)}{d\gamma} \frac{dl_p(\gamma=0)}{d\gamma^T} + \frac{d^2 l_p(\gamma=0)}{d\gamma d\gamma^T} \right) \gamma dF(\gamma; \sigma_G^2) \right) \\ &= l_p(\gamma=0) + \log \left(1 + \frac{1}{2} \text{tr} \left(\frac{dl_p(\gamma=0)}{d\gamma} \frac{dl_p(\gamma=0)}{d\gamma^T} + \frac{d^2 l_p(\gamma=0)}{d\gamma d\gamma^T} \right) \sigma_G^2 \right) \\ &= l_p(\gamma=0) + \log \left(1 + \frac{1}{2} \text{tr}(\mathbf{W}_2 \mathbf{G}^T \mathbf{E}(\Delta - \epsilon)(\Delta - \epsilon)^T \mathbf{E} \mathbf{G} \mathbf{W}_2 - \mathbf{W}_2 \mathbf{G}^T \mathbf{E} \mathbf{V} \mathbf{E} \mathbf{G} \mathbf{W}_2) \sigma_G^2 \right) \end{aligned}$$

where we define $\mathbf{V} = \text{diag} \left\{ \int_0^\infty I(U_i \geq t) e^{X_i \beta + G_i W_1 \theta} d\widehat{\Lambda}_0(t) - \left(\frac{e^{X_i \beta + G_i W_1 \theta}}{\sum_{i=1}^n I(U_i \geq t) e^{X_i \beta + G_i W_1 \theta}} \right)^2 \right\}$

$$\frac{\partial \widetilde{l}_p}{\partial \sigma_G^2} = \frac{\frac{1}{2} \text{tr}(\mathbf{W}_2 \mathbf{G}^T \mathbf{E}(\Delta - \epsilon)(\Delta - \epsilon)^T \mathbf{E} \mathbf{G} \mathbf{W}_2 - \mathbf{W}_2 \mathbf{G}^T \mathbf{E} \mathbf{V} \mathbf{E} \mathbf{G} \mathbf{W}_2)}{1 + \frac{1}{2} \text{tr}(\mathbf{W}_2 \mathbf{G}^T \mathbf{E}(\Delta - \epsilon)(\Delta - \epsilon)^T \mathbf{E} \mathbf{G} \mathbf{W}_2 - \mathbf{W}_2 \mathbf{G}^T \mathbf{E} \mathbf{V} \mathbf{E} \mathbf{G} \mathbf{W}_2) \sigma_G^2}$$

Under the null, $\frac{\partial \widetilde{l}_p}{\partial \sigma_G^2} = \frac{1}{2} \text{tr}(\mathbf{W}_2 \mathbf{G}^T \mathbf{E}(\Delta - \epsilon)(\Delta - \epsilon)^T \mathbf{E} \mathbf{G} \mathbf{W}_2 - \mathbf{W}_2 \mathbf{G}^T \mathbf{E} \mathbf{V} \mathbf{E} \mathbf{G} \mathbf{W}_2)$

Using $\text{tr}(\mathbf{A} - \mathbf{B}) = \text{tr}(\mathbf{A}) - \text{tr}(\mathbf{B})$, we can rewrite the partial derivative as the following:

$$\frac{\partial \widetilde{l}_p}{\partial \sigma_G^2} = \frac{1}{2} [\text{tr}(\mathbf{W}_2 \mathbf{G}^T \mathbf{E}(\Delta - \epsilon)(\Delta - \epsilon)^T \mathbf{E} \mathbf{G} \mathbf{W}_2) - \text{tr}(\mathbf{W}_2 \mathbf{G}^T \mathbf{E} \mathbf{V} \mathbf{E} \mathbf{G} \mathbf{W}_2)]$$

Using $\text{tr}(\mathbf{AC}) = \text{tr}(\mathbf{CA})$, the first term becomes $\frac{1}{2} [\text{tr}((\Delta - \epsilon)^T \mathbf{E} \mathbf{G} \mathbf{W}_2 \mathbf{W}_2 \mathbf{G}^T \mathbf{E}(\Delta - \epsilon))]$

The first term of the partial derivative is scalar, and since the trace of a scalar is equal to itself,

$$\frac{\partial \bar{l}_p}{\partial \sigma_G^2} = \frac{1}{2} \left((\Delta - \epsilon)^T E G W_2 W_2 G^T E (\Delta - \epsilon) - \text{tr}(W_2 G^T E V E G W_2) \right)$$

Taking twice the first term, we have the test statistic,

$$Q = (\Delta - \epsilon)^T E G W_2 W_2 G^T E (\Delta - \epsilon)$$

where $Q \sim \sum_{j=1}^q \zeta_j \chi_{1,j}^2$, where ζ_j 's are eigenvalues of the matrix

$$W_2 G^T E (V - V X (X^T V X)^{-1} X^T V) E G W_2$$

3.6 Simulation Studies

3.6.1 Type I Error

3.6.1.1 Type I Error Simulation Settings

To evaluate type 1 error of the proposed approach, we performed several simulation studies where there is genotype main effect but no GE interaction effect in the model. To simulate the genotypes, we used SeqSIMLA software, which can simulate sequence data for unrelated individuals or families with user-specified pedigree structures. We simulated 3000 unrelated individuals and used reference sequence based on 1000 Genomes Project for European populations (Chung and Shih 2013). For each of 10,000 replicates, assuming proportional hazards, we simulated the survival time from a Weibull distribution (Bender, Augustin, and Blettner 2005) with covariates age, sex, and smoking as the environmental variable from:

$$Time = \sqrt{-\frac{4\log(V)}{\exp(0.005(\mathbf{age} - 50) + 0.05\mathbf{sex} + 0.3\mathbf{smoke} + \mathbf{G}\boldsymbol{\theta})}}$$

where V was randomly sampled from standard uniform distribution (0,1), \mathbf{age} was generated from a normal distribution with mean of 50 and standard deviation equal to 5, \mathbf{sex} was generated from a Bernoulli distribution with probability 0.5, and \mathbf{smoke} was generated from a Bernoulli distribution with probability 0.5. We simulated variants where the directions of the main effect (represented by $\boldsymbol{\theta}$) on risk of the minor allele are either all positive or mixed with 50% positive and 50% negative. $\boldsymbol{\theta}$ consists of effect sizes for the causal SNPs and they are determined by

$$\theta_i = \frac{h}{2MAF_i(1 - MAF_i)}$$

where MAF is the minor allele frequency of SNP i and h is a constant calculated as

$$h = \frac{R^2}{\mathbf{v}^T L \mathbf{v}}$$

and R^2 , the proportion of variance explained by the causal SNPs, is fixed at 1%. The correlations between the SNPs are in matrix L , and \mathbf{v} is a vector that indicates the direction of the SNP effects.

We simulated four different censoring schemes for censoring time C : **1**) $C \sim \text{Unif}(0, 2.5)$; **2**) $C \sim \text{Unif}(0, 4)$; **3**) $C \sim \text{Unif}(0, 8)$; **4**) No censoring. In order, these correspond to approximately 60%, 40%, 20%, and 0% censoring, respectively. From the simulated censoring times, we can calculate the event time $U = \min(T, C)$ with the event

indicator $\Delta = I(T \leq C)$. We also varied the proportion of causal variants with MAF less than 5% in the model from 20, 50, to 80%.

To compare the performance of coxGE, we used the rareGE method that can test the association of the GE interaction for rare variants with binary or continuous traits under GLMM framework (Chen, Meigs, and Dupuis 2014). Since logistic regression cannot take into account of censoring, we utilized the event indicator only to apply rareGE without taking into account of the survival time. Specifically, we set the censored observations as controls and those who experienced the event as cases. For the scenario when there is no censoring, we cannot apply rareGE since all the individuals in our dataset will be set to cases.

Under the null hypothesis, we evaluated type I error for coxGE and rareGE at varying asymptotic thresholds, $\alpha = 0.01$ and 0.001 . Furthermore, for the observed anti-conservative scenarios with 40% and 60% censoring for coxGE, we then calibrated the observed p-values to obtain relevant empirical thresholds by generating an empirical distribution of p-value under the null hypothesis by pooling all the asymptotic p-values from the 10,000 replicates. Then we assessed the significance in the power analysis with the empirical thresholds that yield 0.1% in type I error. For the 20% and 0% censoring scenarios where we do not observe type 1 error inflation, we used the asymptotic thresholds $\alpha = 0.01$ and 0.001 .

3.6.1.2 Type I Error Results

Tables 3 and 4 include type 1 error results and empirical thresholds for causal variants where the directions of the effects of the risk alleles are all positive and mixed with 50% positive and 50% negative, respectively. In general, we observe very similar patterns in Tables 3 and 4. For coxGE, we observe inflation in type 1 error evaluated at 1% and 0.1% for the 40% and 60% censoring scenarios. Empirical thresholds to meet type 1 error of 1% decreases slightly as the percent of causal variants in the model increases. When the percent of censoring decreases from 60% to 40%, empirical thresholds to meet type 1 error at 1% increase, meaning the inflation is less severe. When the percent of censoring decreases to even lower levels, inflation becomes less severe and coxGE seems to meet correct type 1 error. From these simulation results, the performance of coxGE seems to stay consistent whether the directions of the risk alleles of the genetic main effects are all positive or mixed.

When rareGE is applied to analyze time-to-event data, we observe that type 1 error is deflated in most censoring scenarios. This is because logistic regression is not capable of properly handling censored observations and take into account of the survival time, which results in loss of information. Therefore, we recommend not to use logistic regression methods to analyze time-to-event outcomes.

The score test in the Cox model can sometimes be anti-conservative when sample size is small (Chen et al. 2015; Fleming, O'Sullivan, and Harrington 1987). To evaluate the

performance of coxGE with smaller samples, we ran simulations with sample size of 1000 (Refer to Appendices D and E for simulation results with sample size of 1000). We see that the inflation in type 1 error becomes even more severe when our sample size decreases.

Table 3: Type 1 error results for rareGE and coxGE and empirical significance threshold for coxGE with genotype main effects where the directions of the risk alleles are all positive (n=3000)

+/-/0	% censor	α (%)	rareGE (%)	coxGE (%)	coxGE_new_threshold (%)	coxGE_empirical_threshold (%)
4/0/16	60	1.00	0.79	1.15	0.81	0.98
		0.10	0.082	0.121	0.09	0.10
	40	1.00	0.72	1.12	0.93	1.00
		0.10	0.088	0.10	0.10	0.10
	20	1.00	0.70	0.99	1.00	0.99
		0.10	0.091	0.094	0.10	0.094
	0	1.00	NA	1.00	1.00	1.00
		0.10	NA	0.102	0.10	0.101
10/0/10	60	1.00	0.83	1.17	0.80	1.00
		0.10	0.076	0.117	0.09	0.10
	40	1.00	0.82	1.14	0.84	1.00
		0.10	0.082	0.112	0.092	0.10
	20	1.00	0.80	1.01	1.00	1.01
		0.10	0.08	0.10	0.10	0.10
	0	1.00	NA	0.97	1.00	0.97
		0.10	NA	0.097	0.10	0.097
16/0/4	60	1.00	0.86	1.21	0.77	0.99
		0.10	0.074	0.12	0.092	0.10
	40	1.00	0.84	1.12	0.98	1.00
		0.10	0.08	0.115	0.088	0.10
	20	1.00	0.64	1.02	1.00	1.02
		0.10	0.072	0.102	0.10	0.102
	0	1.00	NA	1.03	1.00	1.01
		0.10	NA	0.10	0.10	0.10

+/-/0: number of variants whose main genotype effects are positive/ negative/ neutral
 α is the type 1 error level

NA in rareGE: unable to apply rareGE since all the individuals are set as cases

coxGE_new_threshold is the empirical threshold for 40% and 60% censoring

coxGE_empirical_threshold is the coxGE type I error assessed at coxGE_new_threshold

Table 4: Type 1 error results for rareGE and coxGE and empirical significance threshold for coxGE with genotype main effects where the directions of the risk alleles are 50% positive and 50% negative (n=3000)

+/-/0	% censor	α (%)	rareGE (%)	coxGE (%)	coxGE_new_threshold (%)	coxGE_empirical_threshold (%)
2/2/16	60	1.00	0.79	1.16	0.84	1.00
		0.10	0.086	0.119	0.091	0.10
	40	1.00	0.74	1.10	0.96	0.99
		0.10	0.084	0.102	0.096	0.10
	20	1.00	0.72	1.03	1.00	1.03
		0.10	0.10	0.101	0.10	0.101
	0	1.00	NA	1.02	1.00	1.02
		0.10	NA	0.102	0.10	0.10
5/5/10	60	1.00	0.81	1.18	0.81	0.98
		0.10	0.079	0.122	0.094	0.10
	40	1.00	0.81	1.12	0.90	1.00
		0.10	0.08	0.114	0.09	0.10
	20	1.00	0.75	1.01	1.00	1.01
		0.10	0.081	0.102	0.10	0.102
	0	1.00	NA	1.00	1.00	1.01
		0.10	NA	0.098	0.10	0.10
8/8/4	60	1.00	0.81	1.22	0.80	1.00
		0.10	0.077	0.116	0.093	0.10
	40	1.00	0.77	1.14	0.92	0.99
		0.10	0.081	0.113	0.094	0.10
	20	1.00	0.67	1.02	1.00	1.02
		0.10	0.068	0.097	0.10	0.097
	0	1.00	NA	1.00	1.00	1.00
		0.10	NA	0.101	0.10	0.102

+/-/0: number of variants whose main genotype effects are positive/ negative/ neutral
 α is the type 1 error level

NA in rareGE: unable to apply rareGE since all the individuals are set as cases

coxGE_new_threshold is the empirical threshold for 40% and 60% censoring

coxGE_empirical_threshold is the coxGE type I error assessed at coxGE_new_threshold

3.6.2 Power

3.6.2.1 Power Simulation Settings

To assess power, we simulated data under the alternative hypothesis, where we include a gene by smoking interaction effect in addition to the genotype main effects. Similar to the type I error simulation, genotypes were simulated in the same manner using the SeqSIMLA software. For each scenario, we simulated 10,000 phenotype datasets from

$$Time = \sqrt{\frac{4\log(V)}{\exp(0.005(\mathbf{age} - 50) + 0.05\mathbf{sex} + 0.3\mathbf{smoke} + \mathbf{G}\boldsymbol{\theta} + \mathbf{smoke} * \mathbf{G}\boldsymbol{\gamma})}}$$

where V , \mathbf{age} , \mathbf{sex} , and \mathbf{smoke} were generated from the same distribution described in the type I error simulation study and the genotype effects $\boldsymbol{\theta}$ were determined the same way as in our null simulation study. We varied the censoring schemes and proportion of causal variants included in the model in the same way we defined them in type 1 error settings. We considered the scenarios where the directions of the interaction effect of a causal variant (represented by $\boldsymbol{\gamma}$) are either the same or opposite to the directions of the corresponding genetic main effect (represented by $\boldsymbol{\theta}$). The magnitude of interaction effect $\boldsymbol{\gamma}$ were determined by

$$\gamma_i = \sqrt{\frac{h}{2MAF_i(1 - MAF_i)Var(smoke)}}$$

where MAF_i is the minor allele frequency of SNP i and h is a constant for all causal SNPs calculated as

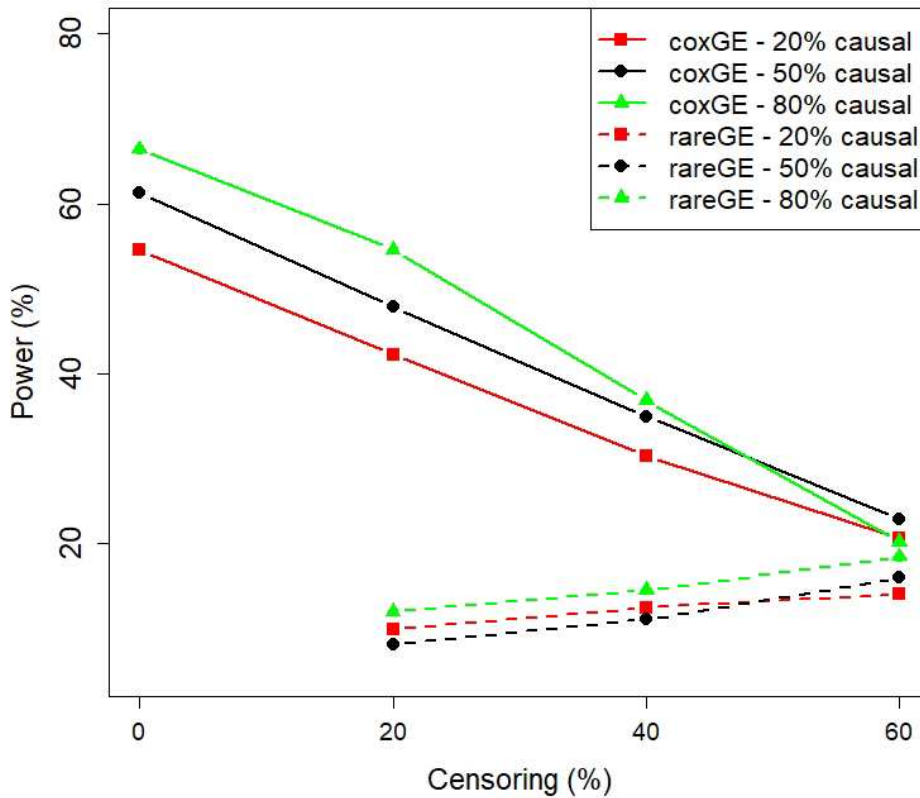
$$h = \frac{R^2}{v^T L v}$$

and R^2 , the proportion of variance explained by gene by smoking interaction, is fixed at 1%. The correlations between the SNPs are in matrix L , and v is a vector that indicates the direction of the interaction effects. For coxGE, we evaluated power under the empirical thresholds that meet 0.1% in type 1 error rates shown in Tables 3 and 4. For rareGE, we evaluated power under asymptotic significance level of $\alpha = 0.001$.

3.6.2.2 Power Results

Power simulation results, where the directions of the risk alleles for genetic main effects and GE interactions are all positive, are presented in Figure 3. For coxGE, when the % of censoring is higher, especially for the 60% scenario, we see a dramatic loss in power. Not surprisingly, we achieve fairly moderate power when there is no censoring at all. When the proportion of causal variants in the model increases, we see that power generally increases. For rareGE, we observe that power increases slightly as the % of censoring (proportion of controls) increases, but in general, power for the rareGE is extremely low across all censoring scenarios compared to that for the coxGE. (Refer to Appendix F for the results where the directions of the risk alleles for genetic main effects and GE interactions are opposite).

Figure 3: Power for coxGE and rareGE with 20%, 50%, and 80% causal variants where the directions of the risk alleles for genetic main effects and GE interactions are all positive



coxGE is evaluated at the empirical threshold that yield 0.1% in type 1 error
 rareGE is evaluated at the asymptotic threshold $\alpha = 0.001$

3.7 Application to the Framingham Heart Osteoporosis Data

We apply our method to the Framingham Osteoporosis Study bone fracture data to test gene-based interaction on time-to-fracture. We consider smoking as our environmental

variable, which has been shown to be associated with higher risk of bone fracture (Al-Bashaireh et al. 2018; Ward 2016; Law and Hackshaw 1997; Lorentzon et al. 2007).

Bone tissue changes throughout life. After age 20-30 years, when peak bone mass is achieved, the skeleton undergoes “remodeling” to replace older bone with newly formed bone. This continuous remodeling process can result in bone loss because of age related imbalance in the resorption of old bone and the formation of new bone. The loss of bone can become severe enough to be called “osteoporosis” with an increase in the risk of fractures (Dimitriou and Giannoudis 2013). Bone fractures can lead to pain, long-term disability, death, and considerable economic costs. As there is evidence for genetic contribution to bone mineral density (BMD), it would be useful to determine the genetic and environmental factors associated with osteoporosis to design the best combination of treatments to prevent fractures (Sosa et al. 2014).

There is now increasing evidence that cigarette smoking is a risk factor for osteoporotic fracture, BMD, and loss of bone mass (Al-Bashaireh et al. 2018; Ward 2016; Law and Hackshaw 1997; Lorentzon et al. 2007). Therefore, it is important to investigate and uncover potential gene x smoking interactions to better understand the interplay of our genes and the detrimental effects of smoking on the skeletal system. We evaluate whether there are modification effects of smoking status on genetic risk for time-to-fracture outcomes. We consider genes that are nearest to the loci previously found to be significantly associated with BMD from a large scale meta-analysis on lumbar spine and

femoral neck BMD to test their interaction with smoking in our analysis (Estrada et al. 2012).

We analyzed the genotype data from the Illumina V1.0 Exome Chip and select variants with MAF less than 5% and those that are annotated as either stop-gain/loss, splice, or missense to minimize noise. We started with 56 genes previously reported to be significantly associated with BMD, but only 15 genes were considered in our analysis since the rest were filtered due to MAF and/or annotation (Estrada et al. 2012). We set the baseline to clinical examination 6 (spanning from 1995-1998), where the participants were under the dual energy X-ray absorptiometry scan to measure their BMD. The participants were then followed until February 15th 2019. We excluded the following fractures from the analysis: toes, fingers, skull, or facial fractures. We selected unrelated individuals from the Offspring Cohort using PC-Air, which results in 618 independent samples that are included in the gene x smoking interaction analysis (Conomos, Miller, and Thornton 2016). In our model, we adjusted for age, sex, BMI, and smoking main effect. There are 280 males (45.3% males), 503 non-smokers (81.4% non-smokers), and their age ranges from 41 to 85 (median = 65). The mean observed survival time is 9.45 years with standard deviation of 5.53 years.

Table 5 summarizes the analysis results for the 15 genes we considered in our analysis. With our proposed method, we found two genes, *SPTBN1* (p-value = 1.62E-3) and *CDKALI* (p-value = 1.82E-3), that are significantly interacting with smoking on time-to-fracture at $\alpha = 0.05/15 = 0.003$.

Table 5: Association results of gene by smoking interaction on time-to-fracture in the Framingham Osteoporosis Study

Gene	Chromosome	# of variants	coxGE p-value
<i>ZBTB40</i>	1	8	0.794
<i>SPTBN1</i>	2	6	1.62E-3
<i>KIAA2018</i>	3	9	0.096
<i>MEPE</i>	4	4	0.565
<i>CDKAL1</i>	6	2	1.82E-3
<i>SUPT3H</i>	6	2	0.629
<i>XKR9</i>	8	4	0.064
<i>CPN1</i>	10	8	0.295
<i>DCDC5</i>	11	6	0.613
<i>ERC1</i>	12	5	0.300
<i>AKAP11</i>	13	13	0.136
<i>AXINI</i>	16	6	0.073
<i>SMG6</i>	17	9	0.364
<i>C17orf53</i>	17	10	0.335
<i>GPATCH1</i>	19	6	0.156

3.8 Discussion

In this chapter, we propose an approach called coxGE to detect GE interaction for time-to-event outcomes. Under this model, we treat the main effects of the genetic variants as fixed effects and use ridge regression to shrink the parameters to prevent potential multicollinearity. We set the GE interaction effects as random and implement a variance component score test. Our simulation studies show that when the percent of censoring is high, we see an inflation in type 1 error and loss of power. When there is an increase in percent of causal variants in the model, we see an increase in power. When the percent of

censoring is low or when there is no censoring, coxGE maintains correct type 1 error and moderate power. Two genes, *SPTBN1* and *CDKALI*, show statistical significance in interacting with smoking on time-to-fracture in the analysis of data from the Framingham Heart Study.

Our real data illustration for the proposed approach identify *SPTBN1* and *CDKALI* to be significantly modified the association between smoking and time-to-fracture although a larger, independent set of subjects is needed to confirm these findings. The *SPTBN1* gene is predicted to be a causal gene associated with BMD in a bone co-expression network analysis (Riaz et al. 2016). In gene expression analysis by Hu et al., *SPTBN1* gene is found to be differentially expressed in the biological network implicated in lung carcinogenesis from never-smoker and current smoker patients (Hu and Chen 2015). To our knowledge, this gene has not been assessed in a gene x smoking interaction study in BMD, but the findings from the two aforementioned studies and the evidence from coxGE potentially suggest rare variants in *SPTBN1* gene interacting with smoking may affect time-to-fracture. *CDKALI* is another gene that came up significant in our analysis. This gene is previously reported to be associated with BMI and type 2 diabetes (Tian et al. 2019; Uma Jyothi and Reddy 2015; Okada et al. 2012). Even though no direct association with smoking has been reported with this gene, there is some evidence that cigarette smoking is a risk factor for type 2 diabetes (Maddatu, Anderson-Baucum, and Evans-Molina 2017; Spijkerman et al. 2014). The relationship between BMI and fracture risk is somewhat controversial. Higher BMI is associated with lower fracture risk but lower BMI is a risk factor for fracture

(Johansson et al. 2014; Nielson et al. 2011). Although this gene is not directly associated with BMD, the relationship between BMD, osteoporosis, and type 2 diabetes has been examined in several studies. It was shown that type 2 diabetics have higher risk of fracture compared to non-diabetics, but paradoxically, they also have higher BMD (Ma et al. 2012; Valderrábano and Linares 2018). This may suggest that there may be alterations to the bone for the diabetic patients.

We note that when the sample size is small, the score test in the Cox model is anti-conservative, as shown in Appendix D and E, so it must be used with caution when the effective sample size is too small. Chen et al. showed that using the likelihood ratio test statistic in the Cox model performs better than the score test when the sample size is small (Chen et al. 2015). For future work, we propose to use the likelihood ratio test as an alternative test statistic to improve the performance of coxGE with small sample size.

As mentioned earlier, one downside of using a ridge penalty is that it cannot force some of the regression coefficients to be exactly 0, leaving all the predictors in the model. Consequently, it is incapable of performing variable selection. This property can potentially be problematic if the number of variants included in the model is large. Variable selection is especially important and desirable in high dimensional setting in order to reduce overfitting and remove redundant variables that do not add any information. For future work, we will explore different penalizations to achieve variable selection, which could potentially improve type 1 error and power.

Currently, there are no existing methods in the literature to test GE interaction for rare variants for survival outcomes. Our proposed method, coxGE, aims to address this gap. Given the importance of time-to-event outcomes in medical studies, we believe our proposed method can be a significant contribution in genetic association studies as well as time-to-event analyses.

Chapter 4 Exploring Regularization Methods in Detecting Gene by Environment Interaction in Rare Variant Analysis for Time-to-Event Outcomes

4.1 Introduction

Rapid evolution of next generation sequencing has significantly increased throughput, but this benefit also comes with a number of methodological challenges with the analysis of large data sets. Penalized regression has received much attention in genetic and genomic analyses due to its ability to handle high dimensional data, deal with problems of collinearity, conduct simultaneous variable selection and parameter estimation (Austin, Pan, and Shen 2013; Zhou et al. 2011). Methods such as ridge (Hoerl and Kennard 1970), least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), and elastic net (Zou and Hastie 2005), are deemed promising in these settings. Generally, these methods are able to encourage fitting a parsimonious model comprising of a small number of covariates that are most important.

As mentioned in the previous chapter, ridge regression has one main disadvantage: it cannot reduce the number of explanatory variables in the model. Although it shrinks some of the regression coefficients very close to 0, none of them are dropped from the model. Therefore, we want to explore different regularization methods that will perform variable selection to see if we can resolve the inflation in type 1 error and higher power than we obtained with a ridge penalty.

In this chapter, we will implement and compare the model performance of two additional penalties, LASSO and elastic net, to the ridge penalty we imposed in coxGE in Chapter 3. LASSO and elastic net are both capable of shrinking some parameters to exactly 0 and thereby permits feature selection. Each of these methods has different strengths and weaknesses so it is important to examine their performance under various simulation scenarios.

This chapter is organized as follows. In sections 4.2 and 4.3, we will introduce LASSO and elastic net penalties and define the coxGE statistics respectively. In section 4.4, we conduct simulations under various settings and compare the results to the coxGE with ridge penalty from Chapter 3. Finally, we illustrate both approaches in testing gene by smoking interaction on time-to-fractures in section 4.5 and conclude with summary and future work in section 4.6.

4.2 LASSO

LASSO regression adds a L1-norm penalty term, which is the absolute value of the magnitude of the coefficients, to the loss function (Tibshirani 1996). Unlike ridge regression, this regularization forces the variables with minor contribution to the model to be exactly 0, effectively selecting a subset of variables to remain in the model. The loss function with LASSO penalty is defined as:

$$L_{lasso}(\beta) = \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

The parameter λ controls the strength of shrinkage. If $\lambda = 0$, all variables will be retained in the model and it will be reduced to standard ordinary least squares and as $\lambda \rightarrow \infty$, all coefficient estimates will move towards 0. Therefore, LASSO has an advantage over ridge penalty since LASSO is capable of removing some features and performing variable selection and thereby reducing the complexity of the model to prevent overfitting. In genetics, only a modest number of variants out of the total number of variants is suspected to be causal. Applying the LASSO penalty seems like an ideal way to exclude variants with minimal effects and to keep only those that are relevant in the model.

In the case when the number of predictors is greater than the number of samples ($p > n$), LASSO is only able to select at most n variables before the model saturates. In this case, LASSO is not a favorable variable selection method. Another disadvantage of LASSO is that when there are highly correlated variables, LASSO tends to arbitrarily select only one variable from the group. For example, in SNP or gene expression data, we would like to automatically include a highly correlated group of variables all together, while discarding the trivial variants. However, LASSO is not able to retain a group of highly correlated variables in the model, a property that is not ideal. To date, various extensions of LASSO have been proposed, such as elastic net (Zou and Hastie 2005), adaptive LASSO (Zou 2006), fused LASSO (Tibshirani et al. 2005), and group LASSO (Yuan and Lin 2006).

4.2.1 LASSO in coxGE

Using the notations defined in Chapter 3, we can write the log partial likelihood with the LASSO penalty as the following:

$$l_L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) - \lambda \sum_{j=1}^p |\theta_j|$$

Calculations will show that the partial derivatives are:

$$\frac{\partial l_L}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \mathbf{M}$$

$$\frac{\partial l_L}{\partial \boldsymbol{\theta}} = \mathbf{W} \mathbf{G}^T \mathbf{M} + \lambda \text{sign}(\theta_j), \text{ where } \text{sign}(\theta_j) = \begin{cases} 1, & \theta_j > 0 \\ 0, & \theta_j = 0 \\ -1, & \theta_j < 0 \end{cases}$$

$$\frac{\partial l_L}{\partial \boldsymbol{\gamma}} = \mathbf{W} \mathbf{G}^T \mathbf{E} \mathbf{M}$$

Due to the change in the likelihood, LASSO regularization will affect the Martingale residual calculation and therefore the Q statistic. Following the derivations of coxGE presented in Chapter 3, we can obtain the Q statistic as

$$\mathbf{Q}_{lasso} = (\boldsymbol{\Delta} - \boldsymbol{\epsilon})^T \mathbf{E} \mathbf{G} \mathbf{W}_2 \mathbf{W}_2 \mathbf{G}^T \mathbf{E} (\boldsymbol{\Delta} - \boldsymbol{\epsilon})$$

where $\mathbf{Q}_{lasso} \sim \sum_{j=1}^q \zeta_j \chi_{1,j}^2$, and ζ_j 's are eigenvalues of the matrix

$$\mathbf{W}_2 \mathbf{G}^T \mathbf{E} (\mathbf{V} - \mathbf{V} \mathbf{X} (\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}) \mathbf{E} \mathbf{G} \mathbf{W}_2$$

4.3 Elastic Net

Elastic net was introduced to address the potential issue of unstable variable selection and the limitation on the number of selected variables in LASSO (Zou and Hastie 2005).

Elastic net combines the ridge and LASSO penalties to obtain a hybrid behavior of L1 and L2 regularizations to encourage groups of highly correlated variables to be selected together in the model and to perform variable selection. The loss function with elastic net penalty is defined as:

$$L_{elastic\ net}(\beta) = \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \left(\frac{(1 - \kappa)}{2} \sum_{j=1}^p \beta_j^2 + \kappa \sum_{j=1}^p |\beta_j| \right)$$

where κ is a mixing parameter, which is restricted between 0 and 1, balancing ridge and LASSO penalties. κ is selected to match the desired balance of variable selection and coefficient shrinkage. When $\kappa = 0$, the penalty function will reduce to ridge term and if $\kappa = 1$, we are left with just the LASSO term. Therefore, by choosing the appropriate κ between 0 and 1, we can achieve the benefits of both ridge and LASSO. Because we have two tuning parameters to approximate, elastic net will be more computationally intensive than ridge and LASSO.

4.3.1 Elastic Net in coxGE

Using the same notations defined in Chapter 3, we can write the log partial likelihood for the elastic net penalty as the following:

$$l_{en}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) - \lambda \left(\frac{(1 - \kappa)}{2} \sum_{j=1}^p \theta_j^2 + \kappa \sum_{j=1}^p |\theta_j| \right)$$

Calculations will show that the partial derivatives are:

$$\frac{\partial l_{en}}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \mathbf{M}$$

$$\frac{\partial l_{en}}{\partial \boldsymbol{\theta}} = \mathbf{W} \mathbf{G}^T \mathbf{M} + \lambda(1 - \kappa)\boldsymbol{\theta}_j + \lambda \kappa \text{sign}(\boldsymbol{\theta}_j), \text{ where } \text{sign}(\boldsymbol{\theta}_j) = \begin{cases} 1, & \boldsymbol{\theta}_j > 0 \\ 0, & \boldsymbol{\theta}_j = 0 \\ -1, & \boldsymbol{\theta}_j < 0 \end{cases}$$

$$\frac{\partial l_{en}}{\partial \boldsymbol{\gamma}} = \mathbf{W} \mathbf{G}^T \mathbf{E} \mathbf{M}$$

Elastic net regularization will affect the Martingale residual calculation due to the change in the likelihood and therefore the Q statistic. Following the derivations of coxGE presented in Chapter 3, we can obtain the Q statistic as

$$\mathbf{Q}_{enet} = (\boldsymbol{\Delta} - \boldsymbol{\epsilon})^T \mathbf{E} \mathbf{G} \mathbf{W}_2 \mathbf{W}_2 \mathbf{G}^T \mathbf{E} (\boldsymbol{\Delta} - \boldsymbol{\epsilon})$$

where $\mathbf{Q}_{enet} \sim \sum_{j=1}^q \zeta_j \chi_{1,j}^2$, where ζ_j 's are eigenvalues of the matrix

$$\mathbf{W}_2 \mathbf{G}^T \mathbf{E} (\mathbf{V} - \mathbf{V} \mathbf{X} (\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}) \mathbf{E} \mathbf{G} \mathbf{W}_2$$

4.4 Simulation

4.4.1 Type I Error

4.4.1.1 Type I Error Simulation Settings

To evaluate type 1 error, we performed several simulation studies where there are genotype main effect but no GE interaction effect. We used the same set of genotypes that we simulated in Chapter 3 and chose a region that spans from 1,100 base pairs to 1,900 base pairs in chromosome 1. For each of 10,000 replicates, assuming proportional hazards, we simulated the survival time from a Weibull distribution (Bender, Augustin, and Blettner 2005) with covariates age, sex, and smoking as the environmental variable from:

$$Time = \sqrt{-\frac{4\log(V)}{\exp(0.005(\mathbf{age} - 50) + 0.05\mathbf{sex} + 0.3\mathbf{smoke} + \mathbf{G}\boldsymbol{\theta})}}$$

where V was randomly sampled from uniform distribution with mean 0 and standard deviation of 1, \mathbf{age} was generated from a normal distribution with mean of 50 and standard deviation equal to 5, \mathbf{sex} was generated from a Bernoulli distribution with probability 0.5, and \mathbf{smoke} was generated from a Bernoulli distribution with probability 0.5. $\boldsymbol{\theta}$ consist of effect sizes for the causal SNPs and they are determined by:

$$\theta_i = \frac{h}{2MAF_i(1 - MAF_i)}$$

where MAF_i is the minor allele frequency of SNP i and h is a constant calculated as

$$h = \frac{R^2}{v^T L v}$$

where R^2 , the proportion of variance explained by the causal SNPs, is fixed at 1%. The correlations between the SNPs are in matrix L , and v is a vector that indicates the direction of the SNP effects.

We simulated four different censoring schemes for censoring time C : **1**) $C \sim \text{Unif}(0, 2.5)$; **2**) $C \sim \text{Unif}(0, 4)$; **3**) $C \sim \text{Unif}(0, 8)$; **4**) No censoring. From this, we can calculate the event time as $U = \min(T, C)$ with the event indicator $\Delta = I(T \leq C)$. We also varied the proportion of causal variants with MAF less than 5% in the model from 20 to 50 to 80%. In previous simulations shown in Chapter 3, the correlation between the variants was overall very low, where the majority of pair-wise correlation range from -0.005 to 0.05. For the observed anti-conservative scenarios with 40% and 60% censoring for all three methods, we calibrated the observed p-values to obtain the relevant empirical thresholds by generating an empirical distribution of p-value under the null hypothesis by pooling all the asymptotic p-values from the 10,000 replicates.

In addition to this low LD setting, we also performed simulation study where we selected causal variants in the model with high LD structure, generally with pair-wise correlation greater than 0.5 (See Appendix I and J for the correlation structures for 20% and 50% causal variant setting). Since ridge, LASSO, and elastic net penalties are known to behave differently in the presence of highly correlated variables, it would be relevant to know if there is one method that outperforms in certain scenarios. Similar to the low LD setting, for the ridge method, we used the empirical thresholds that yield approximately

1% and 0.1% in type 1 error since there was an inflation in type 1 error for 40% and 60% censoring scenarios. We also observed an inflation in type 1 error for 60% censoring scenario for both α levels for the elastic net penalty, so we also used the empirical thresholds. LASSO only showed inflation in the 60% censoring setting when evaluated at $\alpha = 0.01$, so we used the empirical threshold for this setting only for evaluating type 1 error.

4.4.1.2 Type I Error Results for Interaction Test

Tables 6 includes the results for type 1 error from 10,000 simulation replicates for low LD setting evaluated at asymptotic thresholds $\alpha = 0.01$ and 0.001 (refer to Appendix G for type 1 error results evaluated at the empirical thresholds). In the 20% causal variant scenario, the performance of ridge, LASSO, and elastic net do not differ very much in the 40% and 60% censoring scenarios. When the % of censoring decreases, we see that LASSO and elastic net are deflated at $\alpha = 0.001$, while ridge regression meets the correct type 1 error level. We see more deflation for LASSO and ridge as the percent of censoring decreases in both 50% and 80% causal variant scenarios. Between the two methods, LASSO is a bit more conservative than elastic net when the percent of censoring is low at $\alpha = 0.001$. Generally, when the LD between the variants is low, LASSO and elastic net penalties are conservative at low alpha levels, while ridge meets correct type 1 error when there is no censoring or when the % of censoring is low.

We refer to Table 7 for high LD setting results evaluated at asymptotic thresholds $\alpha = 0.01$ and 0.001 (refer to Appendix H for type 1 error results evaluated at the empirical thresholds). In the 20% causal variant scenario, we see that both LASSO and elastic net show deflation, while ridge regression meets correct type 1 error when the percent of censoring is low. In general, we observe more severe deflation of type 1 error in LASSO than in elastic net. Ridge regression shows similar patterns to the low LD setting in the presence of highly correlated causal variants in the model, but LASSO and elastic net penalties show deflation in most scenarios.

Table 6: Type 1 error results for interaction using coxGE with ridge, LASSO, and elastic net penalties for low LD setting at asymptotic thresholds $\alpha = 0.01$ and 0.001

+/-/0	% Censored	α (%)	Ridge (%)	LASSO (%)	Elastic net (%)
4/0/16	60	1	1.15	1.23	1.20
		0.1	0.121	0.124	0.126
	40	1	1.12	1.08	1.16
		0.1	0.10	0.112	0.116
	20	1	0.99	0.95	0.89
		0.1	0.094	0.09	0.098
10/0/10	60	1	1.17	1.27	1.28
		0.1	0.117	0.121	0.124
	40	1	1.14	1.05	0.98
		0.1	0.112	0.10	0.099
	20	1	1.01	1.00	0.82
		0.1	0.10	0.095	0.097
16/0/4	60	1	1.21	1.30	1.30
		0.1	0.12	0.124	0.125
	40	1	1.12	0.82	0.81
		0.1	0.115	0.112	0.113
	20	1	1.02	0.97	0.98
		0.1	0.102	0.091	0.095
0	1	1.03	0.94	0.86	
	0.1	0.10	0.084	0.089	

+/-/0: number of variants with main genotype effect sizes that are positive, negative, and neutral

α is the asymptotic threshold

Ridge is the type 1 error evaluated under α

LASSO is the type 1 error evaluated under α

Elastic net is the type 1 evaluated under α

Table 7: Type 1 error results for interaction using coxGE with ridge, LASSO, and elastic net penalties for high LD setting at asymptotic thresholds $\alpha = 0.01$ and 0.001

+/-/0	% Censored	α (%)	Ridge (%)	LASSO (%)	Elastic net (%)
4/0/16	60	1	1.26	1.33	1.27
		0.1	0.118	0.083	0.123
	40	1	1.20	0.71	0.92
		0.1	0.112	0.077	0.09
	20	1	1.04	0.72	0.80
		0.1	0.103	0.08	0.086
10/0/10	60	1	0.98	0.72	0.78
		0.1	0.094	0.065	0.08
	40	1	1.21	1.23	1.12
		0.1	0.116	0.10	0.112
	20	1	1.12	0.90	0.94
		0.1	0.11	0.082	0.09
16/0/4	60	1	1.05	0.71	0.91
		0.1	0.103	0.062	0.08
	40	1	1.02	0.78	0.84
		0.1	0.096	0.065	0.082
	20	1	1.29	1.21	1.38
		0.1	0.117	0.08	0.13
16/0/4	40	1	1.08	0.89	0.95
		0.1	0.11	0.075	0.096
	20	1	1.03	0.80	0.91
		0.1	0.10	0.071	0.09
	0	1	0.99	0.76	0.83
		0.1	0.098	0.068	0.08

+/-/0: number of variants with main genotype effect sizes that are positive, negative, and neutral

α is the asymptotic threshold

Ridge is the type 1 error evaluated under α

LASSO is the type 1 error evaluated under α

Elastic net is the type 1 evaluated under α

4.4.2 Power

4.4.2.1 Power Simulation Settings

To assess power, we simulated data under the alternative hypothesis, where we included a gene by smoking interaction effect in addition to the genotype main effects. Similar to the type I error simulation, genotypes were simulated in the same manner using the SeqSIMLA software. We simulated 10,000 phenotype datasets from

$$Time = \sqrt{\frac{4\log(V)}{\exp(0.005(\mathbf{age} - 50) + 0.05\mathbf{sex} + 0.3\mathbf{smoke} + \mathbf{G}\boldsymbol{\theta} + \mathbf{smoke} * \mathbf{G}\boldsymbol{\gamma})}}$$

where V , \mathbf{age} , \mathbf{sex} , and \mathbf{smoke} were generated from the same distribution described in the type I error simulation study and the genotype effects $\boldsymbol{\theta}$ were determined the same way as in our null simulation study. We varied the censoring schemes and proportion of causal variants included in the model in the same way as we defined them in type 1 error settings. We considered the scenarios where the directions of the interaction effect of a causal variant (represented by $\boldsymbol{\gamma}$) are either in the same or opposite to the directions of the corresponding genetic main effect (represented by $\boldsymbol{\theta}$). The magnitude of the interaction effect $\boldsymbol{\gamma}$ were determined by

$$\gamma_i = \sqrt{\frac{h}{2MAF_i(1 - MAF_i)Var(smoke)}}$$

where MAF_i is the minor allele frequency of SNP i and h is a constant for all causal SNPs calculated as

$$h = \frac{R^2}{v^T L v}$$

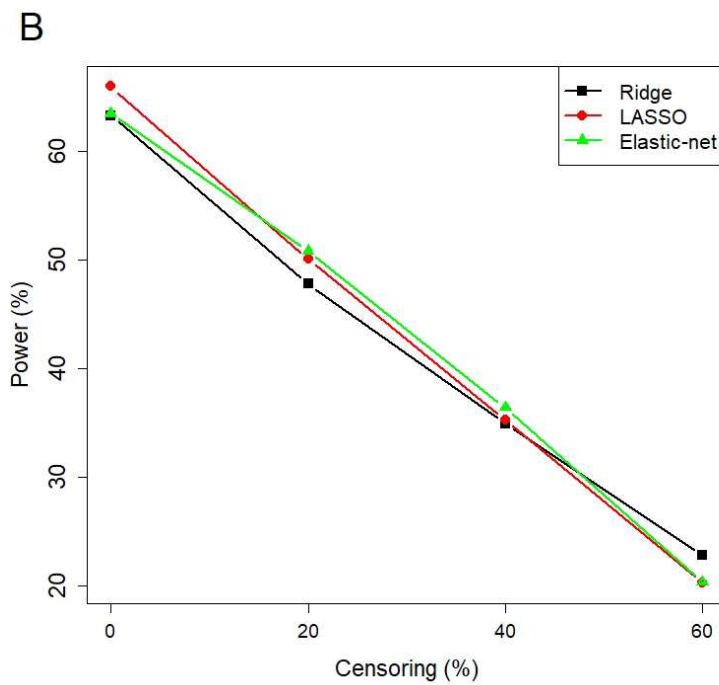
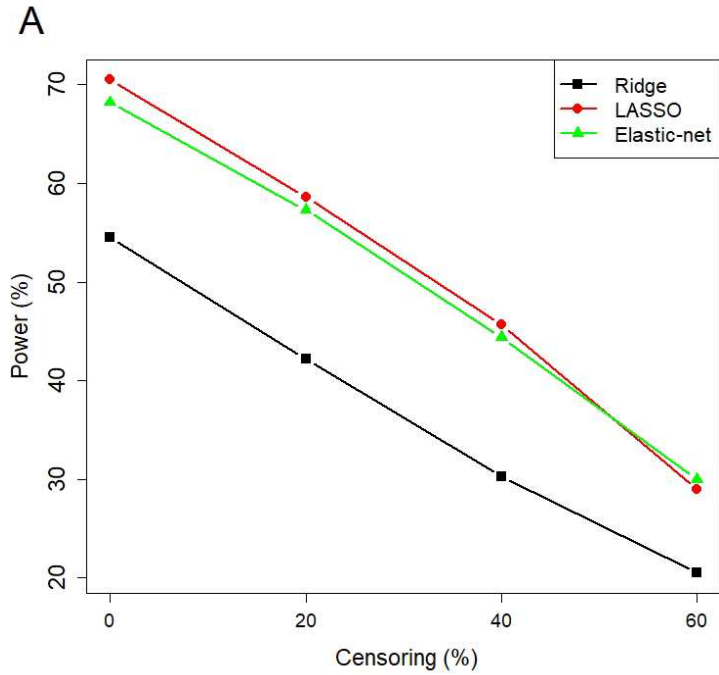
and R^2 , the proportion of variance explained by gene by smoking interaction, is fixed at 1%. The correlations between the SNPs are in matrix L , and v is a vector that indicates the direction of the interaction effects.

4.4.2.2 Power Results

Figures 4 and 5 show power simulation results that are evaluated at the empirical thresholds to achieve 0.1% in type 1 error for low LD and high LD settings, where coxGE with ridge, LASSO, and elastic net penalties are represented in black, red, and green respectively. In the low LD setting, when the percent of censoring increases, especially for 60%, we see a dramatic loss in power in all three approaches because our effective sample size is smaller. Not surprisingly, we achieve fairly moderate power when there is no censoring at all. For the scenario with 20% causal variants, ridge penalty has the lowest power compared to the other two penalties, while LASSO has slightly higher power than elastic net, but this difference is very minimal. When the proportion of causal variants is 50%, all three penalties show very similar power. When the proportion of causal variants is 80%, ridge regression has the highest power, while LASSO has the lowest power. In general, as the proportion of causal variants in the model increases, power for LASSO and elastic net decrease, whereas power for ridge increases except for the 60% censoring scenario.

If we refer to Figure 5 for the high LD settings, in the 20% causal variant scenario, the ridge penalty shows much lower power compared to LASSO and elastic net, which has the highest power, regardless of the percent of censoring. In the 50% causal variant scenario, we come to the same conclusion as the 20% causal variant scenario. We notice that power for LASSO and elastic net decrease, whereas power for ridge regression increases. In the 80% causal variant scenario, we see that ridge regression gains substantial power and has the highest power of all censoring scenarios. In the high LD setting, we observe that the power for elastic net is always greater than that of LASSO.

Figure 4: Power comparison of coxGE with ridge, LASSO, and elastic net penalties with 20% (A), 50% (B), and 80% (C) causal variants in the model in low LD setting



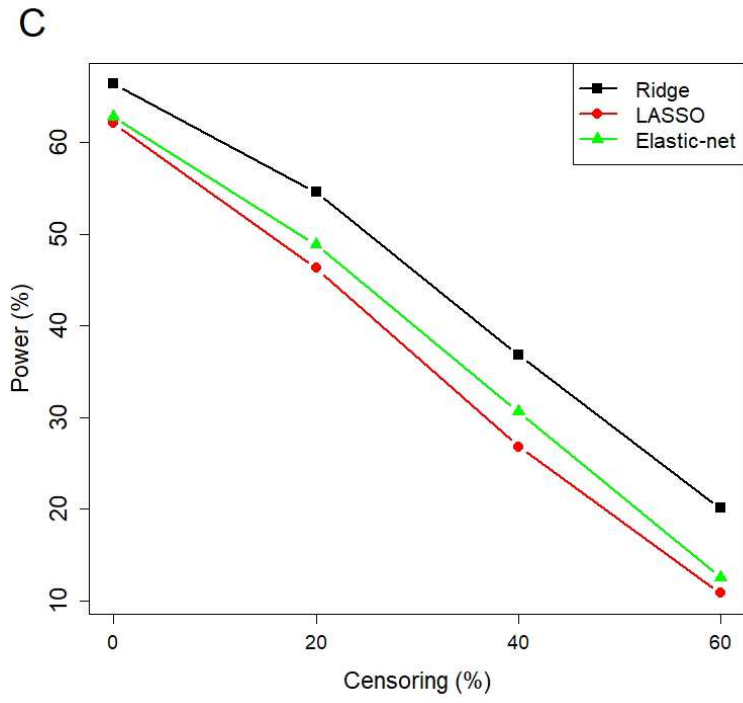
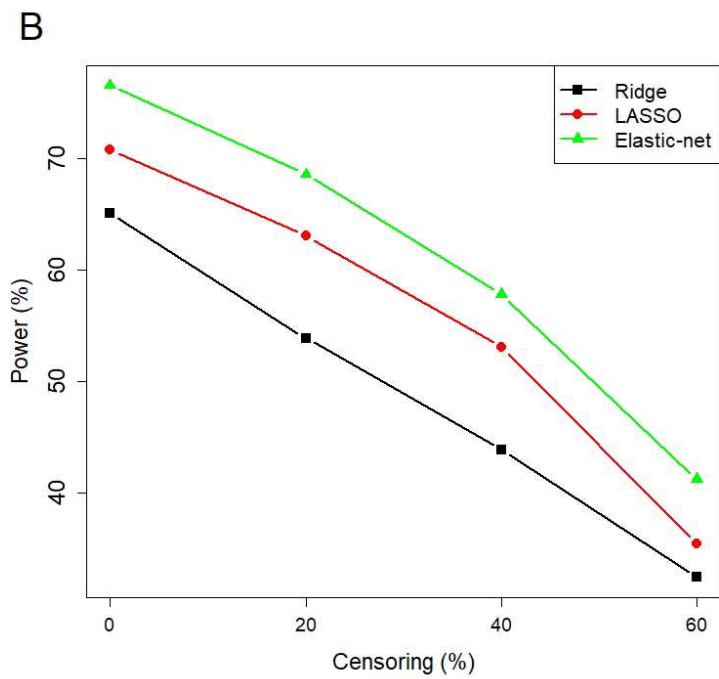
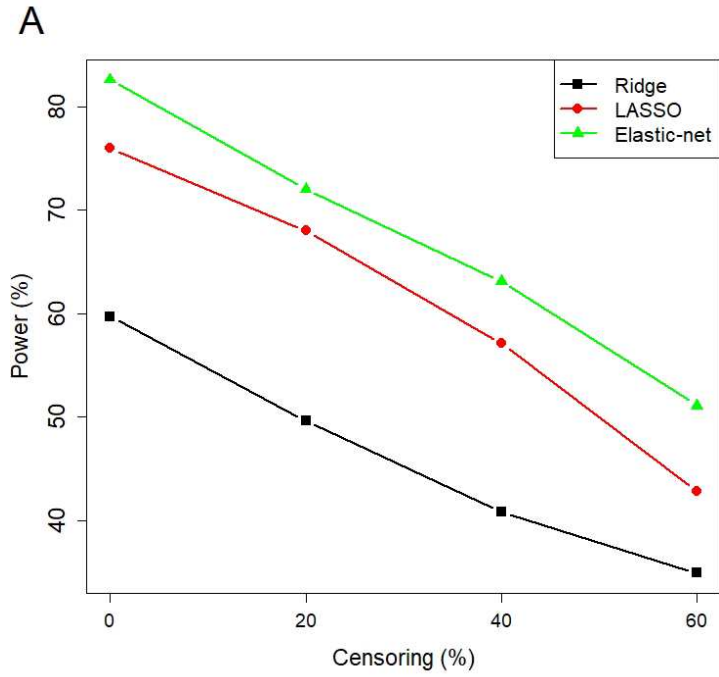
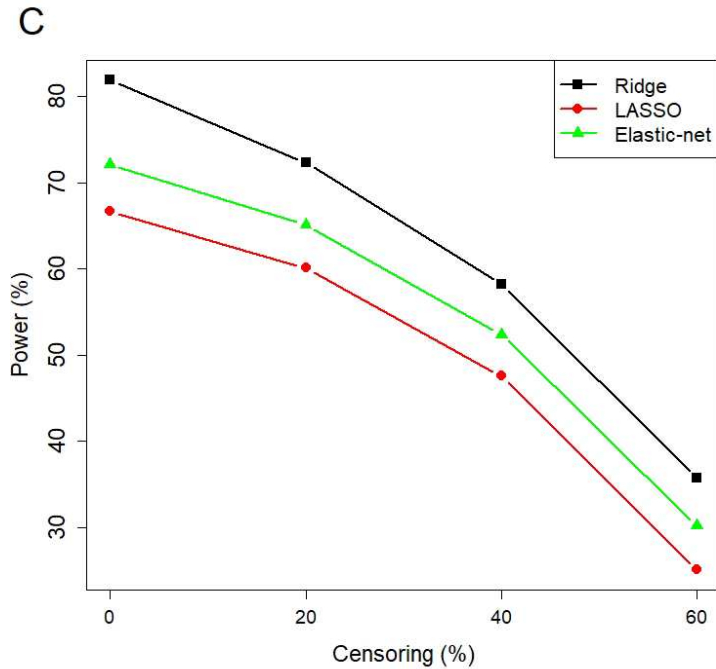


Figure 5: Power comparison of coxGE with ridge, LASSO, and elastic net penalties with 20% (A), 50% (B), and 80% (C) causal variants in the model in high LD setting





4.4.3 Comparison of Number of Variants

In this section, we examine the performance of LASSO and elastic net by comparing the percentage of causal variants that were correctly kept in the model in high LD settings.

We exclude ridge for comparison since it will retain all the variables in the model. Table 8 shows the percentage of causal variants that were correctly kept in the model in high LD setting using LASSO and elastic net from type 1 error and power simulations presented in Tables 6 and 7. As percentage of censoring decreases, the percentage of causal variants that are kept in the model increases. As the proportion of causal variants in the model increases, we see a dramatic decrease in percentage of causal variants retained. We also see that elastic net keeps higher percentage of causal variants in the

model than LASSO does. This is expected since we know that LASSO shows inconsistent variable selection in the presence of highly correlated data, while elastic net aims to address the erratic behavior of LASSO.

Table 8: Percentage of causal variants that were kept in the model in high LD setting

		Results from type 1 error simulation		Results from power simulation	
Scenarios (+/-/0)	% censored	LASSO	Elastic net	LASSO	Elastic net
4/0/16	60	38.2	61.0	33.1	79.1
	40	42.9	73.0	37.3	90.8
	20	50.0	79.2	44.9	92.4
	0	54.1	84.5	52.2	96.9
10/0/10	60	4.95	30.6	2.71	60.3
	40	9.50	43.8	3.10	70.3
	20	12.6	50.6	4.74	74.3
	0	19.6	59.2	6.52	76.5
16/0/4	60	1.31	9.73	1.18	40.4
	40	1.82	18.8	0.50	57.5
	20	2.61	27.1	1.22	64.3
	0	5.20	32.4	1.48	69.2

4.5 Application to the Framingham Heart Osteoporosis Data

Using the same dataset as in Chapter 3, we apply coxGE with LASSO and elastic net penalties to analyze gene x smoking interactions for time-to-fracture. As shown in Table 9, many genes do not have results because all the variants in the gene region were dropped from the model. In Table 10, we show the number of variants that were included in each gene for each method. There are several instances where LASSO and elastic net

retain the same number of variants in the model, and p-values for both approaches are very similar for *CDKALI*, *DCDC5*, *SMG6*, and *AXINI* genes. With the LASSO and elastic net approaches, we find *CDKALI* (LASSO p-value = 8.57E-8, elastic net p-value = 2.58E-7) and *SPTBN1* (p-value = 0.001) to be significant only with the elastic net penalty at $\alpha = 0.003$, a gene also found to be significant using the ridge approach.

Table 9: Association results of gene by smoking interaction on time-to-fracture with ridge, LASSO, and elastic net approaches

Gene	Chromosome	Ridge p-value	LASSO p-value	Elastic net p-value
<i>ZBTB40</i>	1	0.794	NA	NA
<i>SPTBN1</i>	2	1.62E-3	NA	0.001
<i>KIAA2018</i>	3	0.096	0.572	NA
<i>MEPE</i>	4	0.565	NA	0.711
<i>CDKALI</i>	6	1.82E-3	8.57E-8	2.58E-7
<i>SUPT3H</i>	6	0.629	0.756	0.816
<i>XKR9</i>	8	0.064	NA	NA
<i>CPN1</i>	10	0.295	0.206	NA
<i>DCDC5</i>	11	0.613	0.832	0.892
<i>ERC1</i>	12	0.300	NA	NA
<i>AKAP11</i>	13	0.136	0.087	0.109
<i>AXINI</i>	16	0.073	0.299	0.304
<i>SMG6</i>	17	0.364	0.082	0.091
<i>C17orf53</i>	17	0.335	NA	NA
<i>GPATCH1</i>	19	0.156	NA	NA

Table 10: Number of variants included in each gene for the association of gene by smoking interaction on time-to-fracture with ridge, LASSO, and elastic net penalties

Gene	# of variants included		
	Ridge	LASSO	Elastic net
<i>ZBTB40</i>	8	0	0
<i>SPTBN1</i>	6	0	3
<i>KIAA2018</i>	9	4	0
<i>MEPE</i>	4	0	3
<i>CDKAL1</i>	2	2	2
<i>SUPT3H</i>	2	2	2
<i>XKR9</i>	4	0	0
<i>CPN1</i>	8	2	0
<i>DCDC5</i>	6	3	3
<i>ERC1</i>	5	0	0
<i>AKAP11</i>	13	12	10
<i>AXIN1</i>	6	6	6
<i>SMG6</i>	9	7	7
<i>C17orf53</i>	10	0	0
<i>GPATCH1</i>	6	0	0

4.6 Discussion

In this chapter, we investigate two additional penalties, LASSO and elastic net, and compare their performance using coxGE with ridge penalty for detecting gene by environment interaction for time-to-event outcomes. Type 1 error simulations show that LASSO and elastic net methods are conservative at lower alpha levels when the percent of censoring is low in the low LD setting and we observe even more severe deflation in type 1 error in the presence of highly correlated variants. Compared to the LASSO and

elastic net penalty results, the performance of the ridge penalty does not change dramatically between low LD and high LD settings.

As we saw from the real data application, another disadvantage of using LASSO and elastic net is that it can potentially drop all the variants from the model, thus not producing any result. Usually, the tuning parameter λ will be determined by cross validation to choose an optimal λ , as is the case in coxGE, but it is possible to handpick this value if we want to purposely retain or drop more variables from the model. However, this approach is not advised since the goal of cross validation is to select a good value of λ by utilizing the observed data.

For using LASSO and elastic net penalties in coxGE, our underlying hypothesis is different from coxGE with the ridge penalty. With the ridge penalty, we test for GE interaction for all the variants included in the model, regardless of whether the genotype main effects are significant or not. However, with the LASSO and elastic net penalties, we are choosing only those variants that are retained in the model to test their GE interaction effects. Due to their abilities to perform variable selection, this approach assumes that only those variants with significant genetic main effects will be considered for testing their interaction effects. Thus, these two approaches are built on different assumptions.

Initially, we hypothesized that LASSO and elastic net might outperform ridge since they can both perform variable selection, thereby filtering out some of the noise. In the low

LD setting, the performance of the three methods did not show a large difference when the percent of censoring was high. We saw higher inflation for LASSO and elastic net compared to ridge in the 60% censoring scenario. As the percent of censoring decreases, we saw that LASSO and elastic net were conservative compared to ridge. In the high LD setting, LASSO and elastic net were very conservative while ridge was still able to meet correct type 1 error when the percent of censoring was low or when there was no censoring. As suspected, LASSO showed worse performance than elastic net in the presence of highly correlated variables. Since elastic net combines the best features of ridge and LASSO, we expected this method to outperform the other two, but when we are dealing with sparse data, i.e. rare variants, these two variable selection methods do not seem to perform well. Even though ridge can only shrink the coefficients towards 0 and not perform variable selection, it consistently showed decent performance, regardless of the LD between variants. However, we note that power for LASSO and elastic net are generally higher than the ridge regression when the percentage of causal variants in the model is low, but when the genetic main effects are not strong enough in the model, the variants will be discarded and therefore cannot be tested for GE interactions. This could be potentially problematic since the variants we include in the model could all be dropped, as we saw from the real data application using LASSO and elastic net. While it is unlikely that rare variants are highly correlated, the performance of LASSO and elastic net seem to fluctuate depending on the pair-wise correlation of the variants, whereas ridge regression did not seem to suffer from this problem. Therefore, we would

recommend using the ridge penalty over LASSO and elastic net to test GE interaction of rare variants in time-to-event outcomes.

Chapter 5 Summary and Future Work

5.1 Summary

Over the last decade, GWAS have substantially improved our understanding of the genetic architecture of many complex traits. Although GWAS using common variants have made strides in identifying hundreds of loci contributing to complex diseases, these studies are inherently limited and their results only confer relatively small increments of disease risk. In recent years, rapid advances in whole genome sequencing technology have enabled more complete assessment of low frequency and rare variants, and thus, sparked great interest in assessing the roles of rare genetic variants in various complex diseases. Thus, there is a strong need to develop statistical methods to better understand the mechanisms of complex diseases.

In this dissertation, we aim to address both statistical and computational challenges in detecting GE interaction in rare variant analysis. In Chapter 2, we focus on a GE interaction test for rare variants that can accommodate binary or continuous outcomes and correctly incorporate sample relatedness (famGE). The model allows user-defined kernels to be included as weights for genetic main effects and GE interaction effects, and also allows for a pedigree or empirical kinship matrix to account for relatedness. In Chapter 3, we develop GE interaction test of rare variants for time-to-event outcomes using mixed effects Cox regression with ridge penalization on the genetic main effects to reduce multicollinearity between the variants (coxGE). In Chapter 4, we extend this

method to explore LASSO and elastic net penalties and compare the performance of these approaches with the ridge penalty. These proposed methods can be applied to a wide range of phenotypes and various epidemiological studies, which will aid in elucidating the etiology of complex diseases and eventually lead to developing better diagnostics and targeted prevention methods.

5.2 Future Work

5.2.1 Extension of famGE

Further efforts are warranted along the lines of this work. For example, developing a joint test of both genetic main effects and GE interaction effects in the presence of related individuals may enhance the statistical power for identifying genetic associations. Chen et al. proposed a joint test of main and interaction effects for rare variants, and we may follow a similar framework to develop a joint test in the context of famGE (Chen, Meigs, and Dupuis 2014). Considering both the main and interaction effects can potentially identify important genes that were originally missed.

5.2.2 Extension of coxGE

Unlike famGE framework, where we assume genetic main effects to be random, the genetic main effects in coxGE are fixed. Instead, to avoid potential multicollinearity, we impose a penalty term on the genetic main effects to discourage fitting a complex model. However, it was shown in the simulation studies of rareGE by Chen et al that there is inflation in type 1 error rates when the genotype main effects are fixed. Furthermore,

when assessing power, GE interaction with fixed genotype main effects show slightly lower power compared to that of random genotype main effects (Chen, Meigs, and Dupuis 2014). Therefore, it would be preferable to fit the genetic main effects as random since the number of variants in the testing region can be large. We would expect to see an improved type 1 error and power in coxGE if the genetic main effects are random instead.

As shown from rareGE simulation studies in Chapter 2, when familial correlation is ignored, it leads to inflated type 1 error. When we apply coxGE to correlated data, it will most likely also show inflated type 1 error. This is why we select a subset of unrelated individuals from the whole sample. However, we have shown in our type 1 error simulations (Appendix D and E) that when we decrease our sample size, coxGE suffers from higher type 1 error inflation. It is thus important to retain as many samples as possible, especially for a GE interaction test for rare variants. Our current approach is only applicable to unrelated individuals. Therefore, another future effort is to extend coxGE to account for familial correlation in the model to prevent reducing our sample size.

Appendices

Appendix A: Derivation of the famGE statistic to test interaction

The derivation for chapter 2 is based on the GLMM framework. We consider the following GE interaction model:

$$g(E(Y_i)) = g(\mu_i) = \mathbf{X}_i^T \boldsymbol{\alpha} + \mathbf{G}_i^T \mathbf{W}_1 \boldsymbol{\theta} + E_i \mathbf{G}_i^T \mathbf{W}_2 \boldsymbol{\gamma} + d_i,$$

where $g(\cdot)$ is the link function, $\boldsymbol{\alpha}$ is a $p \times 1$ vector associated with the fixed covariate effects, $\boldsymbol{\theta}$ is a $q \times 1$ vector of random effects for the genetic variants, $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effects for GE interaction, d is a $n \times 1$ vector of family correlation, and \mathbf{W}_1 and \mathbf{W}_2 are $q \times q$ diagonal matrices with weights for genetic main effects and GE interaction effects, respectively. We can rewrite this model in matrix notation as:

$$g(\mathbf{E}(\mathbf{Y})) = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\alpha} + \mathbf{b}_1 + \mathbf{b}_2 + \mathbf{b}_3,$$

where $\mathbf{b}_1 = \mathbf{G}\mathbf{W}_1\boldsymbol{\theta}$, $\mathbf{b}_2 = \mathbf{E}\mathbf{G}\mathbf{W}_2\boldsymbol{\gamma}$, and $\mathbf{b}_3 = \mathbf{d}$ and $\mathbf{b} = (\mathbf{b}_1^T, \mathbf{b}_2^T, \mathbf{b}_3^T)^T \sim N(\mathbf{0}, \mathbf{B})$, where $\mathbf{B}(\boldsymbol{\sigma}^2) = \text{diag}\{\sigma_M^2(\mathbf{G}\mathbf{W}_1\mathbf{W}_1\mathbf{G}^T), \sigma_I^2(\mathbf{E}\mathbf{G}\mathbf{W}_2\mathbf{W}_2\mathbf{G}^T\mathbf{E}), \sigma_G^2\boldsymbol{\psi}\}$

For subject i , the quasi-likelihood given random effects \mathbf{b} is defined by:

$$ql_i(\boldsymbol{\alpha}; \mathbf{b}) = \int_{y_i}^{\mu_i} \frac{(y_i - \mu)}{\phi v(\mu)} d\mu$$

The integrated quasi-likelihood function used to estimate $(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2)$ is

$$L(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2) = e^{ql(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2)} \propto |\mathbf{B}|^{-\frac{1}{2}} \int \exp\left\{\sum_{i=1}^n ql_i(\boldsymbol{\alpha}, \mathbf{b}) - \frac{1}{2} \mathbf{b}^T \mathbf{B}^{-1} \mathbf{b}\right\} d\mathbf{b} \quad (\text{A1})$$

High dimensional integration is required to obtain the likelihood function but it is difficult to calculate and maximize, so we use Laplace's method for integral approximation [Breslow & Clayton]. After applying Laplacian transformation, the log of equation (A1) becomes:

$$ql(\alpha, \sigma^2) = -\frac{1}{2} \log|\mathbf{B}| - \frac{1}{2} \log|f''(\tilde{\mathbf{b}})| + f'(\tilde{\mathbf{b}}), \quad (\text{A2})$$

where $\tilde{\mathbf{b}}$ is the solution to

$$f'(\tilde{\mathbf{b}}) = \frac{\partial ql_i}{\partial \mathbf{b}} = -\frac{(y_i - \mu_i) \mathbf{z}_i}{\phi v(\mu_i) g'(\mu_i)} + \mathbf{B}^{-1} \mathbf{b} = 0 \quad (\text{A3})$$

For canonical link functions, the second partial derivative with respect to \mathbf{b} is equal to

$$f''(\tilde{\mathbf{b}}) = \frac{\partial^2 ql_i}{\partial \mathbf{b} \partial \mathbf{b}^T} = \frac{\mathbf{z}_i \mathbf{z}_i^T}{\phi v(\mu_i) [g'(\mu_i)]^2} + \mathbf{B}^{-1} \approx \mathbf{Z}^T \mathbf{D} \mathbf{Z} + \mathbf{B}^{-1}, \quad (\text{A4})$$

where $\mathbf{D} = \text{diag} \left\{ \frac{1}{\phi v(\mu_i) [g'(\mu_i)]^2} \right\}$

Combining (A3) – (A4), equation (A2) becomes

$$ql(\alpha, \sigma^2) = -\frac{1}{2} \log|\mathbf{I} + \mathbf{Z}^T \mathbf{D} \mathbf{Z} \mathbf{B}| - \sum_{i=1}^n ql_i(\alpha, \tilde{\mathbf{b}}) - \frac{1}{2} \tilde{\mathbf{b}}^T \mathbf{B}^{-1} \tilde{\mathbf{b}}, \quad (\text{A5})$$

where $\tilde{\mathbf{b}} = \tilde{\mathbf{b}}(\alpha, \sigma^2)$ is the solution to

$$\frac{\partial}{\partial \mathbf{b}} \left\{ \sum_{i=1}^n ql_i(\alpha, \tilde{\mathbf{b}}) - \frac{1}{2} \tilde{\mathbf{b}}^T \mathbf{B}^{-1} \tilde{\mathbf{b}} \right\} = 0$$

Defining $\mathbf{\Delta} = \text{diag}\{g'(\mu_i)\}$ and assuming that weight matrix \mathbf{D} varies slowly as a function of the mean (following Breslow and Clayton), we maximize the penalized quasi-likelihood by differentiating with respect to α and \mathbf{b} :

$$\frac{\partial ql}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^n \frac{(y_i - \mu_i) \mathbf{X}_i^T}{\phi v(\mu_i) g'(\mu_i)} = \mathbf{X}^T \mathbf{D} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}) \quad (\text{A6})$$

$$\frac{\partial ql}{\partial \mathbf{b}} = \sum_{i=1}^n \frac{(y_i - \mu_i) \mathbf{Z}_i^T}{\phi v(\mu_i) g'(\mu_i)} - \mathbf{B}^{-1} \mathbf{b} = \mathbf{Z}^T \mathbf{D} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{B}^{-1} \mathbf{b} \quad (\text{A7})$$

Defining the working vector $\mathbf{Y}_0 = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b} + \boldsymbol{\Delta}(\mathbf{Y} - \boldsymbol{\mu})$ and by substituting $\boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu})$ with $\mathbf{Y}_0 - \mathbf{X}\boldsymbol{\alpha} - \mathbf{Z}\mathbf{b}$, the solution to (A6) and (A7)

$$\left\{ \begin{array}{l} \mathbf{X}^T \mathbf{D} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{X}^T \mathbf{D} \mathbf{Y}_0 - \mathbf{X}^T \mathbf{D} \mathbf{X} \boldsymbol{\alpha} - \mathbf{X}^T \mathbf{D} \mathbf{Z} \mathbf{b} = \mathbf{0} \\ \mathbf{Z}^T \mathbf{D} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{B}^{-1} \mathbf{b} = \mathbf{Z}^T \mathbf{D} \mathbf{Y}_0 - \mathbf{Z}^T \mathbf{D} \mathbf{X} \boldsymbol{\alpha} - \mathbf{Z}^T \mathbf{D} \mathbf{X} \mathbf{b} - \mathbf{B}^{-1} \mathbf{b} = \mathbf{0} \end{array} \right\}$$

can be expressed using the Fisher scoring method as an iterative solution to the system of equations

$$\begin{bmatrix} \mathbf{X}^T \mathbf{D} \mathbf{X} & \mathbf{X}^T \mathbf{D} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{D} \mathbf{X} & (\mathbf{Z}^T \mathbf{D} \mathbf{Z} + \mathbf{B}^{-1}) \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{D} \mathbf{Y}_0 \\ \mathbf{Z}^T \mathbf{D} \mathbf{Y}_0 \end{bmatrix}$$

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y}_0$$

$$\hat{\mathbf{b}} = \mathbf{B} \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_0 - \mathbf{X} \hat{\boldsymbol{\alpha}})$$

where $\boldsymbol{\Sigma} = \text{Var}(\mathbf{Y}_0) = \mathbf{D}^{-1} + \mathbf{Z} \mathbf{B} \mathbf{Z}^T$

Following Breslow & Clayton, we ignore the dependence between \mathbf{D} and $\boldsymbol{\sigma}^2$ and replace

$\sum_{i=1}^n ql_i(\boldsymbol{\alpha}; \tilde{\mathbf{b}})$ by the Pearson chi-square statistic $-\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{\phi v(\mu_i)}$. After substituting the

maximized values, equation (A5) becomes:

$$\begin{aligned} ql(\hat{\boldsymbol{\alpha}}, \sigma^2) &= -\frac{1}{2} \log |\mathbf{I} + \mathbf{Z}^T \mathbf{D} \mathbf{Z} \mathbf{B}| - \frac{1}{2} \sum_{i=1}^n \frac{a_i (y_i - \hat{\mu}_i)^2}{\phi v(\hat{\mu}_i)} - \frac{1}{2} \hat{\mathbf{b}}^T \mathbf{B}^{-1} \hat{\mathbf{b}} \quad (\text{B1}) \\ &= -\frac{1}{2} \log |\mathbf{D}| - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{Y}_0 - \mathbf{X} \hat{\boldsymbol{\alpha}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_0 - \mathbf{X} \hat{\boldsymbol{\alpha}}), \end{aligned}$$

To adjust for loss of degrees of freedom from estimating α , the restricted maximum likelihood version is

$$ql_R(\hat{\alpha}, \sigma^2) = C - \frac{1}{2} \log|\Sigma| - \frac{1}{2} \log|X^T \Sigma^{-1} X| - \frac{1}{2} (Y_0 - X\hat{\alpha})^T \Sigma^{-1} (Y_0 - X\hat{\alpha})$$

To derive the score test for $H_0: \sigma_I^2 = 0$, the variance component test uses the score statistic derived by taking the derivative of the PQL with respect to σ_I^2

$$\frac{\partial ql}{\partial \sigma_I^2} = -\frac{1}{2} \text{tr}(\Sigma^{-1} E G W_2 W_2 G^T E) + \frac{1}{2} (Y_0 - X\hat{\alpha})^T \Sigma^{-1} E G W_2 W_2 G^T E \Sigma^{-1} (Y_0 - X\hat{\alpha})$$

where $\Sigma = D^{-1} + \sigma_M^2 (G W_1 W_1 G^T) + \sigma_G^2 \psi + \sigma_I^2 (E G W_2 W_2 G^T E)$

Replacing α and the covariance matrix Σ^{-1} by their ML/REML estimates under the null and treating the GE interaction matrix $E G W_2$ as fixed, the first term in the score function is fixed and independent of the phenotype. Taking twice the second term, we have the test statistic

$$Q = (Y_0 - X\hat{\alpha})^T \hat{\Sigma}^{-1} E G W_2 W_2 G^T E \hat{\Sigma}^{-1} (Y_0 - X\hat{\alpha})$$

where $Q \sim \sum_{j=1}^q \lambda_j \chi_{1,j}^2$, where λ_j 's are eigenvalues of the matrix

$$W_2 G^T E (\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} X (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1}) E G W_2$$

Appendix B: Testing interaction with genotypes as fixed effects

We can treat the main genotype effects as fixed and derive the variance component test for interaction. We assume the same model and same distributions for \mathbf{y} and \mathbf{d} , but $\boldsymbol{\theta}$ is now fixed. We define the working vector under the null hypothesis $\tilde{\mathbf{Y}}_{fix} = \mathbf{X}\boldsymbol{\alpha} + \tilde{\mathbf{G}}\boldsymbol{\theta} + \mathbf{d} + \boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu})$, with $var(\tilde{\mathbf{Y}}_{fix}) = \boldsymbol{\Sigma}_{fix} = \sigma_G^2\boldsymbol{\psi} + \mathbf{D}^{-1}$. The test statistic for $H_0: \sigma_I^2 = 0$ is

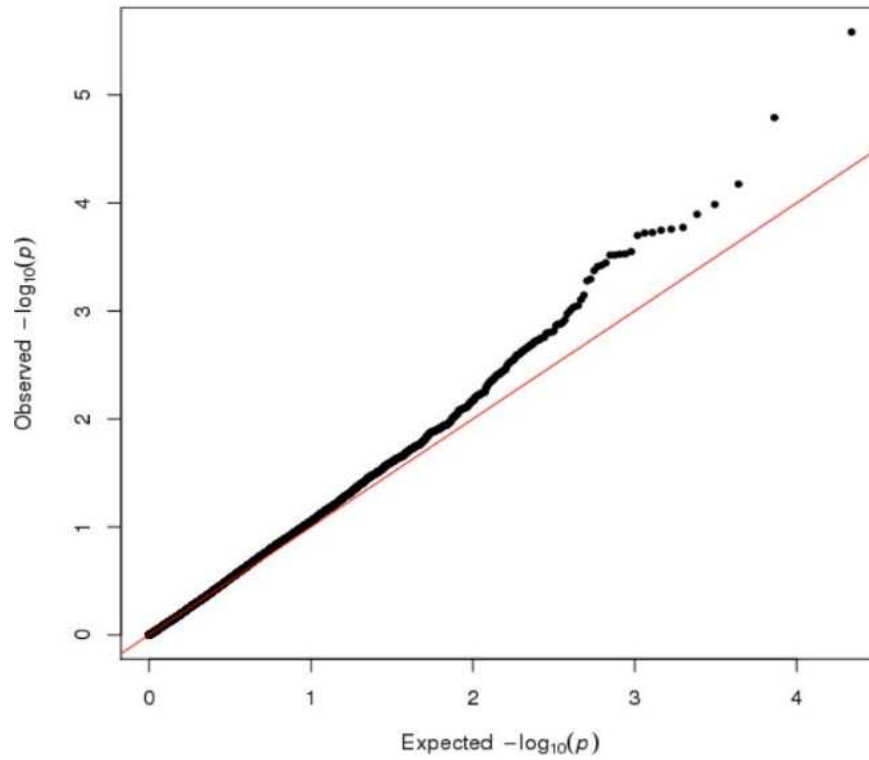
$$\mathbf{Q}_{fix} = (\tilde{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\alpha}} - \tilde{\mathbf{G}}\hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}_{fix}^{-1} \mathbf{E} \mathbf{G} \mathbf{W}_2 \mathbf{W}_2 \mathbf{G}^T \mathbf{E} \hat{\boldsymbol{\Sigma}}_{fix}^{-1} (\tilde{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\alpha}} - \tilde{\mathbf{G}}\hat{\boldsymbol{\theta}})$$

where $\hat{\boldsymbol{\Sigma}}_{fix} = \hat{\sigma}_G^2\boldsymbol{\psi} + \hat{\mathbf{D}}^{-1}$

Under the null hypothesis, $\mathbf{Q}_{fix} \sim \sum_{j=1}^q \lambda_j \chi_{1,j}^2$, where λ_j 's are eigenvalues of the matrix

$$\mathbf{W}_2 \mathbf{G}^T \mathbf{E} (\hat{\boldsymbol{\Sigma}}_{fix}^{-1} - \hat{\boldsymbol{\Sigma}}_{fix}^{-1} \mathbf{X} (\mathbf{X}^T \hat{\boldsymbol{\Sigma}}_{fix}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\boldsymbol{\Sigma}}_{fix}^{-1}) \mathbf{E} \mathbf{G} \mathbf{W}_2$$

Appendix C: QQ plot of famGE applied genome-wide to detect gene x smoking interaction on BMI in the Framingham Heart Study data



Appendix D: Simulation for type 1 errors of coxGE with sample size of 1000 and causal markers all in positive directions

Scenarios (+/-/0)	% censored	α level (%)	coxGE (%)
4/0/16	60	1	1.84
		0.1	0.17
	40	1	1.50
		0.1	0.20
	20	1	1.40
		0.1	0.20
10/0/10	60	1	1.72
		0.1	0.24
	40	1	1.44
		0.1	0.24
	20	1	1.33
		0.1	0.17
16/0/4	60	1	1.64
		0.1	0.29
	40	1	1.79
		0.1	0.19
	20	1	1.36
		0.1	0.14
0	1	1.17	
	0.1	0.12	

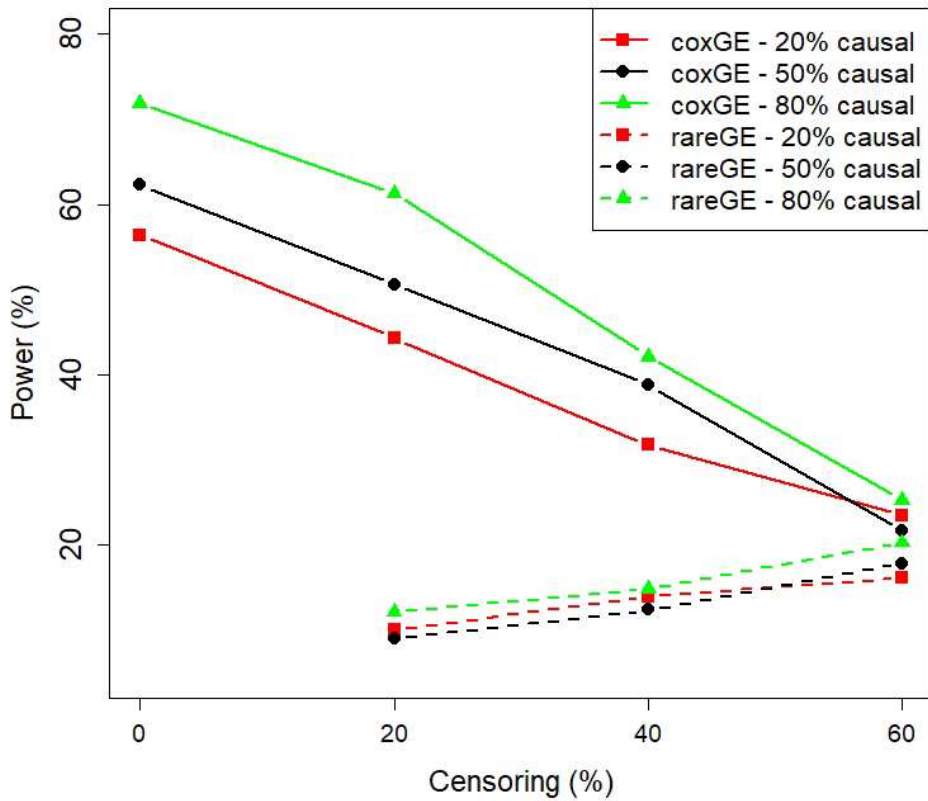
+/-/0: number of variants with main genotype effect sizes that are positive, negative, and neutral

Appendix E: Simulation for type 1 errors for coxGE with sample size of 1000 and causal markers in both positive and negative directions

Scenarios (+/-/0)	% censored	α level (%)	coxGE (%)
2/2/16	60	1	1.82
		0.1	0.24
	40	1	1.58
		0.1	0.18
	20	1	1.44
		0.1	0.14
	0	1	1.08
		0.1	0.12
5/5/10	60	1	1.68
		0.1	0.20
	40	1	1.52
		0.1	0.18
	20	1	1.44
		0.1	0.16
	0	1	1.26
		0.1	0.13
8/8/4	60	1	1.68
		0.1	0.32
	40	1	1.71
		0.1	0.14
	20	1	1.42
		0.1	0.24
	0	1	1.16
		0.1	0.11

+/-/0: number of variants with main genotype effect sizes that are positive, negative, and neutral

Appendix F: Power for coxGE and rareGE with 20%, 50%, and 80% causal variants where the directions of the risk alleles for genetic main effects and GE interactions are opposite



coxGE is evaluated at the empirical threshold that yield $\alpha = 0.001$ in type 1 error
 rareGE is evaluated at the asymptotic threshold $\alpha = 0.001$

Appendix G: Type 1 error results for coxGE with ridge, LASSO, and elastic net penalties and empirical significance threshold for low LD setting

Censor (%)	α (%)	α_{ridge} (%)	Ridge (%)	α_{lasso} (%)	LASSO (%)	α_{enet} (%)	Elastic net (%)
20% causal variant							
60	1.00	0.81	0.98	0.79	1.00	0.90	1.00
	0.10	0.09	0.10	0.094	0.10	0.098	0.10
40	1.00	0.93	1.00	0.95	0.98	0.86	1.00
	0.10	0.10	0.10	0.09	0.10	0.089	0.10
20	1.00	1.00	0.99	1.00	0.95	1.00	0.89
	0.10	0.10	0.094	0.10	0.09	0.10	0.098
0	1.00	1.00	1.00	1.00	0.92	1.00	0.95
	0.10	0.10	0.102	0.10	0.089	0.10	0.092
50% causal variant							
60	1.00	0.80	1.00	0.77	1.00	0.79	0.99
	0.10	0.09	0.10	0.08	0.10	0.096	0.10
40	1.00	0.84	1.00	0.96	1.00	1.14	1.00
	0.10	0.092	0.10	0.10	0.10	0.10	0.10
20	1.00	1.00	1.01	1.00	1.00	1.00	0.82
	0.10	0.10	0.10	0.10	0.095	0.10	0.097
0	1.00	1.00	0.97	1.00	0.96	1.00	0.89
	0.10	0.10	0.097	0.10	0.082	0.10	0.089
80% causal variant							
60	0.99	0.77	0.99	0.75	1.00	0.73	1.00
	0.10	0.092	0.10	0.075	0.10	0.081	0.10
40	1.00	0.98	1.00	1.23	1.00	1.13	1.00
	0.10	0.088	0.10	0.09	0.10	0.092	0.10
20	1.00	1.00	1.02	1.00	0.97	1.00	0.98
	0.10	0.10	0.102	0.10	0.091	0.10	0.095
0	1.00	1.00	1.03	1.00	0.94	1.00	0.86
	0.10	0.10	0.10	0.10	0.084	0.10	0.089

α is the type 1 error level

Ridge is the type 1 error evaluated under empirical threshold, α_{ridge}

LASSO is the type 1 error evaluated under empirical threshold, α_{lasso}

Elastic net is the type 1 error evaluated under empirical threshold, α_{enet}

Appendix H: Type 1 error results for coxGE with ridge, LASSO, and elastic net penalties and empirical significance threshold for high LD setting

Censor (%)	α (%)	α_{ridge} (%)	Ridge (%)	α_{lasso} (%)	LASSO (%)	α_{enet} (%)	Elastic net (%)
20% causal variant							
60	1.00	0.83	1.00	0.78	1.00	0.87	1.00
	0.10	0.09	0.10	0.10	0.083	0.092	0.10
40	1.00	0.89	1.00	1.00	0.71	1.00	0.92
	0.10	0.095	0.10	0.10	0.077	0.10	0.09
20	1.00	1.00	1.04	1.00	0.72	1.00	0.80
	0.10	0.10	0.103	0.10	0.08	0.10	0.086
0	1.00	1.00	0.98	1.00	0.72	1.00	0.78
	0.10	0.10	0.094	0.10	0.065	0.10	0.08
50% causal variant							
60	1.00	0.79	1.00	0.70	0.99	0.79	0.99
	0.10	0.09	0.10	0.10	0.10	0.092	0.10
40	1.00	0.86	1.00	1.00	0.90	1.00	0.94
	0.10	0.091	0.10	0.10	0.082	0.10	0.09
20	1.00	1.00	1.05	1.00	0.71	1.00	0.91
	0.10	0.10	0.103	0.10	0.062	0.10	0.08
0	1.00	1.00	1.02	1.00	0.78	1.00	0.84
	0.10	0.10	0.096	0.10	0.065	0.10	0.082
80% causal variant							
60	1.00	0.77	1.00	0.74	1.00	0.73	1.00
	0.10	0.09	0.10	0.10	0.08	0.081	0.10
40	1.00	0.85	1.00	1.00	0.89	1.00	0.95
	0.10	0.10	0.10	0.10	0.075	0.10	0.096
20	1.00	1.00	1.03	1.00	0.80	1.00	0.91
	0.10	0.10	0.10	0.10	0.071	0.10	0.09
0	1.00	1.00	0.99	1.00	0.76	1.00	0.83
	0.10	0.10	0.098	0.10	0.068	0.10	0.08

α is the type 1 error level

Ridge is the type 1 error evaluated under empirical threshold, α_{ridge}

LASSO is the type 1 error evaluated under empirical threshold, α_{lasso}

Elastic net is the type 1 error evaluated under empirical threshold, α_{enet}

Appendix I: Correlation structure for 20% causal variants in the model

	Snp1	Snp2	Snp3	Snp4
Snp1				
Snp2	0.702			
Snp3	0.922	0.759		
Snp4	0.922	0.759	1.00	

*Only lower triangle of the correlation matrix is displayed

Appendix J: Correlation structure for 50% causal variants in the model

	Snp1	Snp2	Snp3	Snp4	Snp5	Snp6	Snp7	Snp8	Snp9	Snp10
Snp1										
Snp2	0.702									
Snp3	0.922	0.759								
Snp4	0.922	0.759	1.000							
Snp5	-0.033	-0.033	-0.026	-0.026						
Snp6	0.033	0.009	0.036	0.036	-0.037					
Snp7	-0.027	-0.033	-0.024	-0.024	0.678	-0.026				
Snp8	0.053	0.029	0.054	0.054	-0.032	0.844	-0.021			
Snp9	-0.027	-0.033	-0.024	-0.024	0.677	-0.025	1.000	-0.021		
Snp10	-0.027	-0.031	-0.024	-0.023	0.653	-0.024	0.963	-0.021	0.963	

*Only lower triangle of the correlation matrix is displayed

Bibliography

- Al-Bashaireh, Ahmad M. et al. 2018. “The Effect of Tobacco Smoking on Bone Mass: An Overview of Pathophysiologic Mechanisms.” *Journal of Osteoporosis* 2018.
- Allison, P.D. 2011. “Survival Analysis Using SAS : A Practical Guide . Second Edition.” *American Journal of Epidemiology*, 174(4): 503–504.
- Artigas, María Soler et al. 2011. “Genome-Wide Association and Large-Scale Follow up Identifies 16 New Loci Influencing Lung Function.” *Nature Genetics* 43(11): 1082–90.
- Asimit, J L, AG Day-Williams, A P Morris, and E Zeggini. 2012. “ARIEL and AMELI1 A: Testing for an Accumulation of Rare Variants Using next-Generation Sequencing Data.” *Human Heredity* 73(2): 84–94.
- Auer, Paul L, and Guillaume Lettre. 2015. “Rare Variant Association Studies : Considerations , Challenges and Opportunities.” *Genome Medicine* 7(1): 1–11.
- Austin, Erin, Wei Pan, and Xiaotong Shen. 2013. “Penalized Regression and Risk Prediction in Genome-Wide Association Studies.” *Statistical Analysis and Data Mining* 6(4): 315–28.
- Basu, Saonli, and Wei Pan. 2011. “Comparison of Statistical Tests for Disease Association with Rare Variants.” *Genetic Epidemiology* 35(7): 606–19.
- Bender, Ralf, Thomas Augustin, and Maria Blettner. 2005. “Generating Survival Times to Simulate Cox Proportional Hazards Models.” *Stat Med* 1713–1723.
- Breslow, N. E. 1974. “Covariance Analysis of Censored Survival Data.” *Biometrics* 30(1): 89–99.

- Breslow, N. E., and D. G. Clayton. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 88(421): 9–25.
- Cai, Tianxi, Giulia Tonini, and Xihong Lin. 2011. "Kernel Machine Approach to Testing the Significance of Multiple Genetic Markers for Risk Prediction." *Biometrics* 67(3): 975–86.
- Carmen Dering, Claudia Hemmelmann, Elizabeth Pugh, Andreas Ziegler. 2012. "Statistical Analysis of Rare Sequence Variants: An Overview of Collapsing Methods." *Genetic Epidemiology* 35(Suppl 1): 1–11.
- Chasman, Daniel I. et al. 2011. "Genome-Wide Association Study Reveals Three Susceptibility Loci for Common Migraine in the General Population." *Nature Genetics* 43(7): 695–698.
- Chen, Han et al. 2015. "Sequence Kernel Association Test for Survival Traits." *Genetic Epidemiology* 38(3): 191–197.
- Chen, Han, James B. Meigs, and Josée Dupuis. 2013. "Sequence Kernel Association Test for Quantitative Traits in Family Samples." *Genetic Epidemiology* 37(2): 196–204.
- Chen, Han, James B. Meigs, and Josée Dupuis. 2014. "Incorporating Gene-Environment Interaction in Testing for Association with Rare Genetic Variants." *Human Heredity* 78(2): 81–90.
- Chen, Han et al. 2011. "Comparison of Statistical Approaches to Rare Variant Analysis for Quantitative Traits." *BMC Proceedings* 5(SUPPL. 9): S113.
- Chen, Ming-hui, Joseph G Ibrahim, and Qi-man Shao. 2018. "Maximum Likelihood Inference for the Cox Regression Model with Applications to Missing Covariates."

- Journal of Multivariate Analysis* 100(9): 2018–2130.
- Chung, Ren-hua, and Chung-Chin Shih. 2013. “SeqSIMLA : A Sequence and Phenotype Simulation Tool for Complex Disease Studies.” *BMC Bioinformatics*, 14: 199.
- Clark, TG, MJ Bradburn, SB Love, and DG Altman. 2003. “Tutorial Paper Survival Analysis Part I : Basic Concepts and First Analyses.” *British Journal of Cancer* 89(2): 232–238.
- Conomos, Matthew P, Mike Miller, and Timothy Thornton. 2016. “Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness.” *Genetic Epidemiology*, 39(4): 276–293.
- Coombes, Brandon J. 2018. “A Linear Mixed Model Framework for Gene-Based Gene – Environment Interaction Tests in Twin Studies.” *Genetic Epidemiology* 42(7): 648–663.
- Cox, D. 1972. “Regression Models and Life Tables (with Discussion).” *Journal of the Royal Statistical Society*, 34(2): 187–220.
- DeLong, David, George Guirguis, Ying So. 1994. “Efficient Computation of Subset Selection Probabilities with Application to Cox Regression.” *Biometrika* 81(3): 607–611.
- de los Campos, Gustavo, Daniel Sorensen, and Daniel Gianola. 2015. “Genomic Heritability: What Is It?” *PLoS Genetics* 11(5): 1–21.
- Dimitriou, Rozalia, and Peter V Giannoudis. 2013. “The Genetic Profile of Bone Repair.” *Clinical Cases in Mineral and Bone Metabolism*, 10(1): 19–21.
- Efron, Bradley. 1977. “The Efficiency of Cox’s Likelihood Function for Censored Data.”

- Journal of the American Statistical Association* 72(359): 557–565.
- Estrada, Karol et al. 2012. “Genome-Wide Meta-Analysis Identifies 56 Bone Mineral Density Loci and Reveals 14 Loci Associated With Risk of Fracture.” *Nature Genetics* 44(5): 491–501.
- Evangelou, E et al. 2014. “A Meta-Analysis of Genome-Wide Association Studies Identifies Novel Variants Associated with Osteoarthritis of the Hip.” *Annals of the Rheumatic Diseases* 73(12): 2130–2136.
- Fleming, Thomas R., Margaret O’Sullivan, and David P. Harrington. 1987. “Supremum Versions of the Log-Rank and Generalized Wilcoxon Statistics.” *Journal of the American Statistical Association* 82(397): 312–20.
- Flynn, Robert. 2012. “Survival Analysis.” *Journal of Clinical Nursing*: 2789–2797.
- George, Brandon, Samantha Seals, and Inmaculada Aban. 2014. “Survival Analysis and Regression Models.” *Journal of Nuclear Cardiology* 21(4): 686–694.
- Gorski, Mathias et al. 2017. “1000 Genomes-Based Meta-Analysis Identifies 10 Novel Loci for Kidney Function.” *Scientific Reports* 7:45040.
- Hamza, Taye H. et al. 2011. “Genome-Wide Gene-Environment Study Identifies Glutamate Receptor Gene GRIN2A as a Parkinson’s Disease Modifier Gene via Interaction with Coffee.” *PLoS Genetics* 7(8).
- Han, Fang, and Wei Pan. 2010. “A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants.” *Human Heredity* 70(1): 42–54.
- Hancock, Dana B. et al. 2010. “Meta-Analyses of Genome-Wide Association Studies Identify Multiple Loci Associated with Pulmonary Function.” *Nature Genetics*

42(1): 45–52.

- He, Karen Y. et al. 2017. “Rare Variants in Fox-1 Homolog A (RFX1) Are Associated with Lower Blood Pressure.” *PLoS Genetics* 13(3): 1–15.
- He, Meian et al. 2015. “Meta-Analysis of Genome-Wide Association Studies of Adult Height in East Asians Identifies 17 Novel Loci.” *Human Molecular Genetics* 24(6): 1791–1800.
- Hoerl, Arthur E., and Robert W. Kennard. 1970. “Ridge Regression: Applications to Nonorthogonal Problems.” *Technometrics* 12(1): 69–82.
- Hu, Meian et al. 2015. “Meta-Analysis of Genome-Wide Association Studies of Adult Height in East Asians Identifies 17 Novel Loci.” *Human Molecular Genetics* 24(6): 1791–1800.
- Igartua, Catherine, Sahar V. Mozaffari, Dan L. Nicolae, and Carole Ober. 2017. “Rare Non-Coding Variants Are Associated with Plasma Lipid Traits in a Founder Population.” *Scientific Reports* 7(1): 1–13.
- Ionita-Laza, Iuliana et al. 2013. “Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants.” *American Journal of Human Genetics* 92(6): 841–853.
- Jiang, Duo, and Mary Sara McPeck. 2014. “Robust Rare Variant Association Testing for Quantitative Traits in Samples with Related Individuals.” *Genetic Epidemiology* 38(1): 10–20.
- Jiao, Shuo et al. 2013. “SBERIA: Set-Based Gene-Environment Interaction Test for Rare and Common Variants in Complex Diseases.” *Genetic Epidemiology* 37(5): 452–

464.

Johansson, Helena et al. 2014. “A Meta-Analysis of the Association of Fracture Risk and Body Mass Index in Women.” *Journal of Bone and Mineral Research* 29(1): 223–33.

Justice, Anne E. et al. 2017. “Genome-Wide Meta-Analysis of 241,258 Adults Accounting for Smoking Behaviour Identifies Novel Loci for Obesity Traits.” *Nature Communications* 8: 1–19.

Kao, Patrick Y.P. et al. 2017. “Pathway Analysis of Complex Diseases for GWAS, Extending to Consider Rare Variants, Multi-Omics and Interactions.” *Biochimica et Biophysica Acta - General Subjects* 1861(2): 335–353.

Kenneth D. Ward, Robert C. Klesges. 2016. “A Meta-Analysis of the Effects of Cigarette Smoking on Bone Mineral Density.” *Calcified Tissue International* 15(5): 477–91.

Law, M. R., and A. K. Hackshaw. 1997. “A Meta-Analysis of Cigarette Smoking, Bone Mineral Density and Risk of Hip Fracture: Recognition of a Major Effect.” *British Medical Journal* 315(7112): 841–46.

Leclerc, Martin, The Consortium, Jacques Simard, and Lajmi Lakhhal-chaieb. 2015. “SNP Set Association Testing for Survival Outcomes in the Presence of Intrafamilial Correlation.” *Genetic Epidemiology* 39(6): 406-414.

Lee, Seunggeun et al. 2012. “Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies.” *American Journal of Human Genetics* 91(2): 224–37.

Lee, Seunggeun, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. 2014. “Rare-

- Variant Association Analysis: Study Designs and Statistical Tests.” *American Journal of Human Genetics* 95(1): 5–23.
- Lee, Seunggeun, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. 2014. “Rare-Variant Association Analysis: Study Designs and Statistical Tests.” *American Journal of Human Genetics* 95(1): 5–23.
- Li, Bingshan, and Sm Leal. 2008. “Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data.” *The American Journal of Human Genetics* 83: 311–21.
- Liao, Chunxiao et al. 2016. “The Association of Cigarette Smoking and Alcohol Drinking with Body Mass Index: A Cross-Sectional, Population-Based Study among Chinese Adult Male Twins.” *BMC Public Health* 16(1): 1–9.
- Lim, Elise, Han Chen, Josee Dupuis, and Ching-Ti Liu. 2019. “A Unified Method for Rare Variant Analysis of Gene-Environment Interactions.” *Stat Med* 39(6): 801-813
- Lin, Dan Yu, and Zheng Zheng Tang. 2011. “A General Framework for Detecting Disease Associations with Rare Variants in Sequencing Studies.” *American Journal of Human Genetics* 89(3): 354–67.
- Lin, Xihong, Xinyi Lin, Seunggeun Lee, and David C. Christiani. 2013. “Test for Interactions between a Genetic Marker Set and Environment in Generalized Linear Models.” *Biostatistics* 14(4): 667–81.
- Lin, Xinyi et al. 2011. “Kernel Machine SNP-Set Analysis for Censored Survival Outcomes in Genome-Wide Association Studies.” *Genetic Epidemiology* 35(7): 620-631.

- Lin, Xinyi et al. 2016. “Test for Rare Variants by Environment Interactions in Sequencing Association Studies.” *Biometrics* 72(1): 156–164.
- Lorentzon, Mattias, Dan Mellström, Egil Haug, and Claes Ohlsson. 2007. “Smoking Is Associated with Lower Bone Mineral Density and Reduced Cortical Thickness in Young Men.” *Journal of Clinical Endocrinology and Metabolism* 92(2): 497–503.
- Ma, Li, Andrew G. Clark, and Alon Keinan. 2013. “Gene-Based Testing of Interactions in Association Studies of Quantitative Traits.” *PLoS Genetics* 9(2): 1–12.
- Ma, Lili et al. 2012. “Association between Bone Mineral Density and Type 2 Diabetes Mellitus: A Meta-Analysis of Observational Studies.” *European Journal of Epidemiology* 27(5): 319–32.
- Maddatu, Judith, Emily Anderson-Baucum, and Carmella Evans-Molina. 2017. “Smoking and Risk of Type 2 Diabetes.” *Translational Research* 176(1): 101–107.
- Madsen, Bo Eskerod, and Sharon R. Browning. 2009. “A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic.” *PLoS Genetics* 5(2).
- Manolio, Teri A. et al. 2009. “Finding the Missing Heritability of Complex Diseases.” *Nature* 461(7265): 747–53.
- Matsui, Takeshi, and Ian M. Ehrenreich. 2016. “Gene-Environment Interactions in Stress Response Contribute Additively to a Genotype-Environment Interaction.” *PLoS Genetics* 12(7): 1–16.
- Mazo Lopera, Mauricio, Brandon Coombes, and Mariza de Andrade. 2017. “An Efficient Test for Gene-Environment Interaction in Generalized Linear Mixed Models with Family Data.” *International Journal of Environmental Research and Public Health*

14(10): 1134.

- Moreno-Macias, Hortensia, Isabelle Romieu, Stephanie J London, and Nan M Laird. 2010. "Gene-Environment Interaction Tests for Family Studies with Quantitative Phenotypes: A Review and Extension to Longitudinal Measures." *Human Genomics* 4(5): 302–26.
- Morgenthaler, Stephan, and William G. Thilly. 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 - Fundamental and Molecular Mechanisms of Mutagenesis* 615(1–2): 28–56.
- Morris, Andrew P., and Eleftheria Zeggini. 2010. "An Evaluation of Statistical Approaches to Rare Variant Analysis in Genetic Association Studies." *Genetic Epidemiology* 34(2): 188–93.
- Nielson, Carrie M. et al. 2011. "BMI and Fracture Risk in Older Men: The Osteoporotic Fractures in Men Study (MrOS)." *Journal of Bone and Mineral Research* 26(3): 496–502.
- Qi, Wenjing, Andrew S. Allen, and Yi Ju Li. 2019. "Family-Based Association Tests for Rare Variants with Censored Traits." *PLoS ONE* 14(1): 1–17.
- Rask-Andersen, Mathias, Torgny Karlsson, Weronica E. Ek, and Åsa Johansson. 2017. "Gene-Environment Interaction Study for BMI Reveals Interactions between Genetic Factors and Physical Activity, Alcohol Consumption and Socioeconomic Status." *PLoS Genetics* 13(9): 1–20.
- Riancho, José a. 2012. "Genome-Wide Association Studies (GWAS) in Complex

- Diseases: Advantages and Limitations.” *Reumatología clinica* 8(2): 56–57.
- Riaz, Nadeem, Suzanne L Wolden, Daphna Y Gelblum, and J Eric. 2016. “Integrating GWAS and Co-Expression Network Data - Bone.” 118(24): 6072–78.
- Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. “Regularization Paths for Cox Proportional Hazards Model via Coordinate Descent.” *Journal of Statistical Software*, 39(5): 1–13.
- Sosa, Nerea Hernandez De et al. 2014. “Heritability of Bone Mineral Density in a Multivariate Family-Based Study.” *Calcified Tissue International* 94(6): 590–96.
- Spijkerman, Annemieke M.W. et al. 2014. “Smoking and Long-Term Risk of Type 2 Diabetes: The EPIC-InterAct Study in European Populations.” *Diabetes Care* 37(12): 3164–71.
- Stel, Vianda, Friedo Dekker, Giovanni Tripepi, and Carmine Zoccali. 2011. “Survival Analysis II : Cox Regression.” *Nephron. Clinical Practice*, 119(3), c255–260.
- Su, Yu-ru, Chong-zhi Di, and L I Hsu. 2017. “A Unified Powerful Set-Based Test for Sequencing Data Analysis of GxE Interactions.” *Biostatistics* 18(1) 119–131.
- Svishcheva, Gulnara R., Nadezhda M. Belonogova, and Tatiana I. Axenovich. 2014. “FFBSKAT: Fast Family-Based Sequence Kernel Association Test.” *PLoS ONE* 9(6): 1–5.
- Taylor, Amy E. et al. 2014. “Stratification by Smoking Status Reveals an Association of CHRNA5-A3-B4 Genotype with Body Mass Index in Never Smokers.” *PLoS Genetics* 10(12): 1–6.
- Tian, Yanni et al. 2019. “A Novel Polymorphism (Rs35612982) in CDKAL1 Is a Risk

- Factor of Type 2 Diabetes: A Case-Control Study.” *Kidney and Blood Pressure Research* 710061: 1313–26.
- Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Association* 58(1): 267–288.
- Tibshirani, R. et al 2005. “Sparsity and Smoothness via the Fused Lasso.” *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67(1): 91–108.
- Tzeng, Jung Ying et al. 2011. “Studying Gene and Gene-Environment Effects of Uncommon and Common Variants on Continuous Traits: A Marker-Set Approach Using Gene-Trait Similarity Regression.” *American Journal of Human Genetics* 89(2): 277–288.
- Uma Jyothi, Kommoju, and Battini Mohan Reddy. 2015. “Gene-Gene and Gene-Environment Interactions in the Etiology of Type 2 Diabetes Mellitus in the Population of Hyderabad, India.” *Meta Gene* 5: 9–20.
- Valderrábano, Rodrigo J., and Maria I. Linares. 2018. “Diabetes Mellitus and Bone Health: Epidemiology, Etiology and Implications for Fracture Risk Stratification.” *Clinical Diabetes and Endocrinology* 4(1): 1–8.
- Wang, Yuanjia, Yin Hsiu Chen, and Qiong Yang. 2012. “Joint Rare Variant Association Test of the Average and Individual Effects for Sequencing Studies.” *PLoS ONE* 7(3).
- Wu, Michael C. et al. 2011. “Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test.” *American Journal of Human Genetics*

89(1): 82–93.

- Yu, Chenglong et al. 2018. “Low-Frequency and Rare Variants May Contribute to Elucidate the Genetics of Major Depressive Disorder.” *Translational Psychiatry* 8(1).
- Yuan, Ming, and Yi Lin. 2006. “Model Selection and Estimation in Regression with Grouped Variables.” *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 68(1): 49–67.
- Yukinori Okada, Michiaki Kubo, Hiroko Ohmiya, Atsushi Takahash, Natsuhiko et al. 2012. “Common Variants at CDKAL1 and KLF9 Are Associated with Body Mass Index in East Asian Populations.” *Nature Genetics* 23(1): 1–7.
- Zhang, Daowen, and Xihong Lin. 2003. “Hypothesis Testing in Semiparametric Additive Mixed Models.” *Biostatistics* 4(1): 57–74.
- Zhang, Yuan, Shili Lin, and Swati Biswas. 2017. “Detecting Rare and Common Haplotype–Environment Interaction under Uncertainty of Gene–Environment Independence Assumption.” *Biometrics* 73(1): 344–355.
- Zhao, Guolin, Rachel Marceau, Daowen Zhang, and Jung Ying Tzeng. 2015. “Assessing Gene-Environment Interactions for Common and Rare Variants with Binary Traits Using Gene-Trait Similarity Regression.” *Genetics* 199(3): 695–710.
- Zhou, H et al. 2011. “Penalized Regression for Genome-Wide Association Screening of Sequence Data.” *Pac Symp Biocomput.*: 106–17.
- Zou, Hui. 2006. “The Adaptive Lasso and Its Oracle Properties.” *Journal of the American Statistical Association* 101(476): 1418–29.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B* 67(5): 768.

Curriculum Vitae

