Set-based Tests for Genetic Association and Gene-Environment Interaction

by

Zihuai He

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2016

Doctoral Committee:

Professor Bhramar Mukherjee, Co-Chair
Associate Professor Min Zhang, Co-Chair
Professor Sharon Kardia
Assistant Professor Seunggeun Lee

To my parents

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisors Dr. Bhramar Mukherjee and Dr. Min Zhang, for the continuous support of my Ph.D study and related research, for their patience, motivation, and immense knowledge. I could not have imagined having better advisors and mentors for my Ph.D study. I appreciate their vast knowledge and skill in many areas (e.g., statistics, genetics, epidemiology), and their assistance in writing papers and this thesis. I would like to thank the other members of my committee, Dr. Seunggeun Lee, and Dr. Sharon Kardia for the assistance they provided at all levels of the research projects. I would like to thank Dr. Ana Diez Rouz from the Drexel University, and Dr. Jennifer Smith from the department of epidemiology in the University of Michigan for taking time out from their busy schedule to help with the data analysis of the Multi-Ethnic Study of Atherosclerosis.

Very special thanks go out to Dr. Xihong Lin, Dr. Yi Li and Dr. Samuel Kou, without whose motivation and encouragement I would not have considered a graduate career in statistics/biostatistics. They truly made a difference in my life. It was under their tutelage that I developed a focus and became interested in statistics when I was in college. It was though their, persistence, understanding and kindness that I completed my undergraduate degree and was encouraged to apply for graduate training in statistics/biostatistics.

Thanks also goes out to my friends and colleagues: Fan Wu, Zhe Fei, Wenting Cheng, Danting Zhu, Xiaowei Zhan, Dajiang Liu, Hai Shu, Ming Teng, Yebin Tao, Xin Wang, Yumeng Li, Ziqian Geng, Di Yang, Xu Shu, Wei Ding, Shan Kang, Shengchun Kong, Donggeng Xia and Wei Zhao, for their suggestions and advices at times of critical need,

and for all the fun we have had in the last four years.

At last, I would like to thank my family: my parents Jinghua Li and Kan He, for giving birth to me at the first place and supporting me spiritually throughout my life.

TABLE OF CONTENTS

DEDICATION . . . . **ii**

ACKNOWLEDGEMENTS . . . . **iii**

LIST OF FIGURES . . . . **vii**

LIST OF TABLES . . . . **viii**

LIST OF APPENDICES . . . . **xv**

ABSTRACT . . . . **xvi**

CHAPTER

I. Introduction 1

II. Set-based Tests for Genetic Association in Cross-sectional Studies 5

2.1 Introduction 5
2.2 Method 6
2.2.1 Genetic Random Field Model 6
2.2.2 Genetic Random Field Test 10
2.2.3 Robustness to Other Distributions 12
2.3 Application: Dallas Heart Study 13
2.4 Simulation Studies 15

III. Set-based Tests for Genetic Association in Longitudinal Studies 22

3.1 Introduction 22
3.2 Method 24
3.2.1 Longitudinal Genetic Random Field Model 24
3.2.2 Association Test under the Longitudinal Genetic Random Field Model . . 28
3.2.3 Testing for the Joint Effect of Gene and Gene-time Interaction 31
3.3 Application: Multi-Ethnic Study of Atherosclerosis 33
3.4 Simulation Studies 36

IV. Set-Based Tests for Gene-Environment Interaction in Longitudinal Studies 44

4.1 Introduction 44
4.2 Application: Multi-Ethnic Study of Atherosclerosis 47
4.3 Model and Inference 49
4.4 Main Effect Adjustment 54
4.4.1 Gene-environment independence condition 55

4.4.2 Main effect misspecification of E 57
4.4.3 High-dimensionality of G 59
4.5 Numerical Studies 62
4.6 Data Analysis 69

V. Summary and Discussion 74

APPENDICES **79**

BIBLIOGRAPHY **130**

LIST OF FIGURES

FIGURE

C.1 Region MECOM (821 SNPs) in Chinese Americans (775 subjects). The top figure shows its linkage disequilibrium structure, created by Haploview based on HapHap CHD reference samples. Bar plots show the distribution of coefficient squares and eigen-values. The coefficients were estimated by first regressing systolic blood pressure on covariates used in the data analysis of MESA, then regress the residuals on each SNP/PC. 120

C.2 Locus-zoom plots of single SNP analysis of systolic blood pressure. The left panel shows the interaction between perceived healthy food availability and region indexed by rs10850411 in European Americans. The right panel shows the interaction between density of recreational facilities and region CACNB2 in Hispanic Americans. Each dot presents the p-value with respect to one SNP in the region. 125

C.3 The left panel shows the subgroup effect of perceived healthy food availability on systolic blood pressure stratified by the genotype of rs10850411 in European Americans. The right panel shows the effect of density of recreational facilities on systolic blood pressure stratified by the genotype of rs7085587 in Hispanic Americans. 127

C.4 Genome-wide set-based analysis of systolic blood pressure in MESA using the proposed method. Four ethnic groups' p-values were combined using Fisher's method for a meta-analysis after an ethnicity specific analysis. Each dot presents the p-value of one gene. 127

C.5 Genome-wide set-based analysis of diastolic blood pressure in MESA using the proposed method. Four ethnic groups' p-values were combined using Fisher's method for a meta-analysis after an ethnicity specific analysis. Each dot presents the p-value of one gene. 128

C.6 Genome-wide set-based analysis of systolic blood pressure in MESA using iSKAT with the baseline value of repeated measures. Four ethnic groups' p-values were combined using Fisher's method for a meta-analysis after an ethnicity specific analysis. Each dot presents the p-value of one gene. 128

C.7 Genome-wide set-based analysis of diastolic blood pressure in MESA using iSKAT with the baseline value of repeated measures. Four ethnic groups' p-values were combined using Fisher's method for a meta-analysis after an ethnicity specific analysis. Each dot presents the p-value of one gene. 129

LIST OF TABLES

TABLE

2.1 Application to Dallas Heart Study for non-synonymous variants. GenRF: the unweighted genetic random field test; SKAT: the unweighted sequential kernel association test of Wu, et al. (2011); PCR: the princeple component regression test of Guaderman et al. (2007); MinSNP: the MinSNP test considered by Ballard et al. (2010); F-test: the F-test in linear regression; GenRF.w: the genetic random field test with Beta (1, 25) weight of Wu, et al. (2011); SKAT.w: the sequential kernel association test with Beta (1, 25) weight; VT: the variable-threshold test of Price et al. (2010). $*$ indicates p-value is less than or equal to $\alpha = 0.05$. 15

2.2 Type I error rate simulation results under different levels of MAF and sample size (1000 replicates). Each cell contains the type I error rate, i.e., rejection rate when data are generated under the null model. GenRF: the unweighted genetic random field test; SKAT: the unweighted sequential kernel association test of Wu, et al. (2011); GenRF.w: GenRF with Beta(1,25) weight as in Wu, et al. (2011); SKAT.w: SKAT with Beta(1,25) weight. 20

2.3 Power simulation results for common variant analysis under different levels of linkage disequilibrium (LD), interaction effects and causal proportion (1000 replicates). GenRF: the unweighted genetic random field test; SKAT: the unweighted sequential kernel association test of Wu, et al. (2011); PCR: the princeple component regression test of Guaderman et al. (2007); MinSNP: the MinSNP test considered by Ballard et al. (2010); F-test: the F-test in linear regression. 20

2.4 Power simulation results under scenarios involving rare variants with different proportion of causal variants (1000 replicates). Rare Causal Variants: causal variants are rare only; Common & Rare Causal Variants: causal variants are both rare and common; DHS: scenario based on the Dallas Heart Study. GenRF.w: the genetic random field test with Beta (1, 25) weight as in Wu, et al. (2011); SKAT.w: the sequential kernel association test with Beta (1, 25) weight; VT: the variable-threshold test of Price et al. (2010); Other entries as in Table 2. 21

3.1 Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data: top two regions associated with systolic blood pressure/diastolic blood pressure. Each cell shows the p-value. CAU: Caucasians; AFA: African Americans; HIS: Hispanics; CHN: Asians of Chinese descent. Meta: Meta-analysis combining the results of four ethnic groups using Fisher's combined probability test. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is compound symmetric. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome.The column "SNPs" shows the total number of typed and imputed SNPs in each ethnic group. . . . 35

3.2 Type-I Error Rate Corresponding to Different Within-Subject Correlation Structures. Each cell represents the empirical type-I error rate evaluated at α=0.05, 0.01 and 0.001 based on 100000 replicates. The total number of observations is 2,400 and repeated measurements per subject were generated in the same follow-up period according to different correlation structures. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. 40

3.3 Power comparisons when **one** randomly selected SNP is causal and has **a marginal effect**. Each cell represents the empirical power from 500 replicates at level α=0.05. The total number of observations is 2,400 and repeated measurements were recorded in the same follow-up period. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction. 41

3.4 Power comparisons when **one** randomly selected SNP is causal and has **only a gene-time interaction effect**. Each cell represents the empirical power from 500 replicates at level α=0.05. The total number of observations is 2,400 and repeated measurements were recorded in the same follow-up period. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction. 42

3.5 Power comparisons when randomly selected **multiple** SNPs are causal and have **both marginal and interaction effects**. Each cell represents the empirical power from 500 replicates at level α=0.05. The total number of observations is 2,400 and repeated measurements were recorded in the same follow-up period. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction. 43

4.1 Simulation study evaluating the main effect adjustment of E (2526 subjects, 190 SNPs). iSKAT: set based test proposed by Lin et al. (2013). rareGE: rareGE test proposed by Chen et al. (2014) assuming a random main effect of G. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using natural cubic-spline smoothing for E with $\sqrt{m}$ basis functions. GE-true: the proposed test with the correct model, which correctly specifies the main effect of E. The GE methods were implemented using the weighted PCA approach for the main effect of G. Each cell presents type I error rate or power based on 1000 replicates evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07. The calibrated powers with value zero correspond to very high type I errors. 67

4.2 Simulation study evaluating the main effect adjustment of G (775 subjects). A linear main effect of E was fitted. Each cell presents the type I error rate/power based on 1000 replicates. MinP: single SNP analysis using GEE adjusted by the effective number of independent tests (Gao et al., 2008). iSKAT: region based test proposed by Lin et al. (2013). rareGE: rareGE test proposed by Chen et al. (2014) assuming a random main effect of G. GE-none: the proposed test adjusting for none of the SNPs. GE-wPCA/wPCAM-$\sqrt{m}$: the proposed test adjusting for the leading $\sqrt{m}$ components using the weighted PCA and robust(wPCA)/model-based(wPCAM) inference. GE-true: the proposed test with the correct model, which correctly correctly includes all SNPs with non-zero main effects. Type I error rate and power were both evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07. 68

4.3 Simulation study evaluating the use of longitudinal data. GE-CS/AR1: the proposed test with different working correlation (compound symmetric/first order autoregressive). GE-avg.: the proposed test using the average/baseline value of repeated measurements. iSKAT-avg.: cross-sectional iSKAT using the average value of repeated measurements. rareGE-avg.: rareGE test proposed by Chen et al. (2014) using the average value of repeated measurements, and assuming a random main effect of G. The GE methods were implemented using the weighted PCA approach for the main effect of G and natural cubic-spline for the main effect of E adjusting for $\sqrt{m}$ terms. Each cell presents type I error rate or power based on 1000 replicates evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07. 68

4.4 Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data: interactions between neighborhood variables and the region indexed by rs10850411 on systolic blood pressure. Each cell shows the p-value. EUR: European Americans; AFA: African Americans; HIS: Hispanics; CHN: Asians of Chinese descent. Meta: Meta-analysis combining the results of four ethnic groups using Fisher's combined probability test. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E. MinP: minimum p-value test based on GEE. The assumed working correlation is compound symmetric. iSKAT-avg./base.: cross-sectional iSKAT using the average/baseline value of repeated measurements as the outcome. rareGE-avg./base.: cross-sectional rareGE using the average/baseline value of repeated measurements as the outcome, where a random main effect of G is assumed. Bonferroni correction threshold is 0.00043. 73

A.1 Simulation results under different distributions of the response variable (1000 replicates). $*$ indicates results are unavailable due to "sample size is small, need small sample adjustment" and SKAT has no small sample adjustment for IBS kernel. 81

A.2 Simulation results under different levels of heteroscedastic variances (500 replicates). Coefficient: the coefficient of the covariate. $*$ indicates that some predicted means reached 0 or 1. 82

A.3 Dallas Heart Study sequencing data information: number of non-synonymous variants in each gene. MAF: minor allele frequency. 83

B.1 Running times corresponding to different number of total observations. The running times for both fitting the null model and testing the target region C10orf107 (190 SNPs) on a 2.67GHz Linux PC with an Intel Xeon X5650 processor are showed in this table. The numbers of total observations (n) are 3000, 6000 and 10000, approximating the numbers corresponding to CHN, HIS/AFA and CAU ethnic groups respectively, as observed in MESA. CAU: Caucasians; AFA: African Americans; HIS: Hispanics; CHN: Asians of Chinese descent. . . . 93

B.2 Power Studies for Meta/Mega-analysis when Causal Variants are Distinct/Common across Four Ethnic Groups. Each cell represents the empirical power from 500 replicates at level α=0.05. In each ethnic group, one randomly selected SNP is causal and has a marginal effect. LGRF-meta: use LGRF-G to test the region within ethnicity and combine the p-values by Fisher's method. LGRF-mega: jointly tests the four ethnic groups by pooling the individual level data. GEE-meta: use GEE-G to test each SNP within ethnicity and combine the p-values by METAL proposed by Willer, et al. (2010). GEE-mega: test each SNP using the individual level data of four ethnic groups jointly. 94

B.3 Power Studies for evaluating the impact of the genetic similarity metric: Each cell represents the empirical power from 500 replicates at level α=0.05. The number of repeated measurements per subject is six, and there are 400 subjects in each replicate. The correlation structure among repeated measurements is compound symmetric. The parameter configurations are same as the three simulation settings (Tables 3-5) in the main text. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. Both LGRF and LGRF-J use the genetic similarity metric. LGRF-G-IBS: the LGRF test for the marginal effect of a gene using the identity-by state (IBS) similarity. LGRF-J-IBS: the LGRF test for the joint effect of gene and gene-time interaction. The genetic main effect is modeled using IBS similarity. 95

B.4 Power Studies including GenRF: Each cell represents the empirical power from 500 replicates at level α=0.05. The number of repeated measurements per subject is six, and there are 400 subjects in each replicate. The correlation structure among repeated measurements is compound symmetric. Detailed parameters are same as the three simulation settings (Table 3-5) in the main text respectively. LGRF-G: the LGRF test for the marginal effect of a gene using longitudinal dats. LGRF-Avg.: the LGRF test applied to the average of repeated measurements. GenRF-Avg.: the GenRF test applied to the average of repeated measurements. SKAT-Avg.: the SKAT test applied to the average of repeated measurements. 96

B.5 Type-I error rate evaluation at small α level. Each cell represents the empirical type-I error rate of LGRF-G (the LGRF test for the marginal effect of a gene) based on 2.5×10^7 replicates. The total number of observations is 2,400 and repeated measurements per subject were generated in the same follow-up period according to different correlation structures. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. The working correlation assumed in LGRF-G is CS. 97

B.6 Power comparisons when the number of subjects ranges from 1,200 to 6,000 and four repeated measurements were recorded. Randomly selected multiple SNPs are causal and have both marginal and interaction effects. Each cell represents the empirical power from 500 replicates at level α=0.001. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction. 99

B.7 Power comparisons when the number of subjects ranges from 3,600 to 9,600 and four repeated measurements were recorded. Randomly selected multiple SNPs are causal and have both marginal and interaction effects. Each cell represents the empirical power from 500 replicates at level $\alpha = 2.5 \times 10^{-6}$. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction. 100

B.8 Gender distribution of MESA subjects across site and race. Each cell represents the number of subject in the corresponding category. WFU: Wake Forest University, Winston Salem, NC; COL: Columbia University, New York, NY; JHU: Johns Hopkins University, Baltimore, MD; UMN: University of Minnesota, Twin Cities, MN; NWU: Northwestern University, Chicago, IL; UCLA: University of California - Lost Angeles, Los Angeles, CA. 101

B.9 Longitudinal summary of blood pressure phenotypes and covariates we adjusted for in MESA across four exams. Sd: standard deviation. N: number of subject. sBP: systolic blood pressure. dBP: diastolic blood pressure. BMI: body mass index. 101

B.10 Sensitivity analysis of the top 5 principal components (PCs) in MESA. Each cell represents the p-value. The analysis was done by fitting a multivariate linear regression using exam 1 data in MESA. 101

B.11 Chromosomal Region Information for the 29 regions considered in the MESA analysis. 102

B.12 CAU ethnic group sBP (2526 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements. 103

B.13 CAU ethnic group dBP (2526 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements. 103

B.14 AFA ethnic group sBP (1611 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements. 104

B.15 AFA ethnic group dBP (1611 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements. 104

B.16 HIS ethnic group sBP (1449 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements. 105

B.17 HIS ethnic group dBP (1449 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements. 105

B.18 CHN ethnic group sBP (775 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements. 106

B.19 CHN ethnic group dBP (775 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements. 106

B.20 Meta-Analysis of sBP (6361 subjects) in MESA. Four ethnic groups were combined using Fisher's method. . . . 107

B.21 Meta-Analysis of dBP (6361 subjects) in MESA. Four ethnic groups were combined using Fisher's method. . . 107

C.1 Simulation study evaluating the main effect adjustment of G (500 subjects). A linear main effect of E was fitted. Each cell presents the type I error rate/power based on 1000 replicates. MinP: single SNP analysis using GEE adjusted by the effective number of independent tests (Gao et al., 2008). iSKAT: region based test proposed by Lin et al. (2013). rareGE: rareGE test proposed by Chen et al. (2014) assuming a random main effect of G. GE-none: the proposed test adjusting for none of the SNPs. GE-wPCA/wPCAM-$\sqrt{m}$: the proposed test adjusting for the leading $\sqrt{m}$ components using the weighted PCA and robust(wPCA)/model-based(wPCAM) inference. GE-true: the proposed test with the correct model, which correctly correctly includes all SNPs with non-zero main effects. Type I error rate and power were both evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07. 117

C.2 Simulation study evaluating the main effect adjustment of G (775 subjects). A linear main effect of E was included in the working model. Each cell presents the type I error rate/power based on 1000 replicates. GE-PCA/PLS/wPCA-5/$\sqrt{m}$: the proposed test adjusting for the leading 5/$\sqrt{n}$ components using principal component analysis (PCA)/partial least square regression (PLS)/ the weighted PCA (wPCA) approach. GE-true: the proposed test with a correct model. Type I error rate and power were both evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has high type I error rate (> 0.07). The zero calibrated power is due to extremely low p-values when type I error rate is high. 118

C.3 Gender distribution of MESA subjects across site and race. Each cell represents the number of subject in the corresponding category. WFU: Wake Forest University, Winston Salem, NC; COL: Columbia University, New York, NY; JHU: Johns Hopkins University, Baltimore, MD; UMN: University of Minnesota, Twin Cities, MN; NWU: Northwestern University, Chicago, IL; UCLA: University of California - Lost Angeles, Los Angeles, CA. 121

C.4 Marginal association between sBP/dBP and the neighborhood exposures. This is a joint analysis of all four ethnic groups using GEE, adjusting for age, gender, BMI and the socioeconomic status. 121

C.5 Longitudinal summary of blood pressure phenotypes and covariates we adjusted for in MESA across four exams. Sd: standard deviation. N: number of subject. sBP: systolic blood pressure. dBP: diastolic blood pressure. BMI: body mass index. SES: socioeconomic status. DFFS: density of favorable food stores. DRF: density of recreational facilities. PHFA: perceived healthy foods availability. PW: perceived walkability. 121

C.6 Chromosomal Region Information for the 29 regions considered in the MESA analysis. 122

C.7 Marginal association between sBP and the 29 regions in MESA. Four ethnic groups were combined using Fisher's method. MinP: minimum p-value test using GEE. SKAT: the SKAT test using the average value of the outcome. LGRF-G/J: the longitudinal genetic random field model proposed by He et al. (2015). The results are from He et al. (2015). Age, gender, BMI and top 2 PCs were adjusted as covariates. 123

C.8 Marginal association between dBP and the 29 regions in MESA. Four ethnic groups were combined using Fisher's method. MinP: minimum p-value test using GEE. SKAT: the SKAT test using the average value of the outcome. LGRF-G/J: the longitudinal genetic random field model proposed by He et al. (2015). The results are from He et al. (2015). Age, gender, BMI and top 2 PCs were adjusted as covariates. 123

C.9 Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data. Each cell shows the p-value. EUR: European Americans; HIS: Hispanics. DFFS: Density of favorable food stores. DRF: Density of recreational facilities. PHFA: Perceived Healthy Food Availability. PW: Perceived walkability. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E. MinP: minimum p-value test based on GEE. iSKAT-avg./base.: cross-sectional iSKAT using the average/baseline value of repeated measurements as the outcome. Bonferroni correction threshold is 0.00043. . 124

C.10 Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data. Each cell shows the p-value. EUR: European Americans; HIS: Hispanics. DFFS: Density of favorable food stores. DRF: Density of recreational facilities. PHFA: Perceived Healthy Food Availability. PW: Perceived walkability. GE-base: the proposed test using the baseline measurement of both Y and E. GE-cumulative: the proposed test using the cumulative average of previous measurements. GE: the proposed test as noted in the main text, which assumes full covariate conditional mean (Pepe and Anderson, 1994). $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E are used. Bonferroni correction threshold is 0.00043. 124

C.11 Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data: interactions between neighborhood variables and the region CACNB2 on systolic blood pressure. Each cell shows the p-value. EUR: European Americans; AFA: African Americans; HIS: Hispanics; CHN: Asians of Chinese descent. Meta: Meta-analysis combining the results of four ethnic groups using Fisher's combined probability test. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E. MinP: minimum p-value test based on GEE. The working correlation assumed in LGRF is compound symmetric. iSKAT-avg./base.: cross-sectional iSKAT using the average/baseline value of repeated measurements as the outcome. Bonferroni correction threshold is 0.00043. 125

C.12 Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data. Each cell shows the p-value. DFFS: Density of favorable food stores. DRF: Density of recreational facilities. PHFA: Perceived Healthy Food Availability. PW: Perceived walkability. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E. MinP: minimum p-value test based on GEE. iSKAT-avg.: cross-sectional iSKAT using the average value of repeated measurements as the outcome. iSKAT-base.: cross-sectional iSKAT using the baseline value of repeated measurements as the outcome. Bonferroni correction threshold is 0.00043. 126

LIST OF APPENDICES

APPENDIX

A. Supplementary Materials for Chapter II 80

B. Supplementary Materials for Chapter III 84

C. Supplementary Materials for Chapter IV 108

ABSTRACT

Substantial progress has been made in identifying single genetic variants predisposing to common complex diseases. Nonetheless, the genetic etiology of human diseases remains largely unknown. Since these traits are likely influenced by the joint effect of multiple variants in a gene/region, a joint analysis of these variants considering linkage disequilibrium (LD) may help to explain additional phenotypic variation. In this dissertation, we present several set-based tests for genetic association/gene-environment interaction in both cross-sectional and longitudinal studies.

In the first project, we propose a new statistical model based on the random field theory, referred to as a genetic random field model (GenRF), for gene/region based association analysis in a cross-sectional study. Using a pseudo-likelihood approach, a GenRF test for the joint association of multiple genetic variants is developed, which has the following advantages: 1. accommodating complex interactions for improved performance; 2. natural dimension reduction; 3. boosting power in the presence of LD; 4. computationally efficient. Simulation studies are conducted under various scenarios. Compared with the sequence kernel association test (SKAT), as well as other more standard methods, GenRF shows overall comparable performance and better performance in the presence of complex interactions. The method is further illustrated by an application to the Dallas Heart Study.

In the second project, we propose a longitudinal genetic random field model (LGRF), to test the association between a phenotype measured repeatedly during the course of an observational study and a set of genetic variants. Generalized score type tests are developed, which we show are robust to misspecification of within-subject correlation. In addition, a joint test incorporating gene-time interaction is further proposed. Computational advancement is made for scalable implementation of the proposed methods in large-scale genome-wide association studies (GWAS). The proposed methods are evaluated through extensive simulation studies and illustrated using data from the Multi-Ethnic Study of Atherosclerosis (MESA).

In the third project, we propose a generalized score type test for set-based inference for gene-environment interaction with longitudinally measured quantitative traits. The test is robust to misspecification of within subject correlation structure and has enhanced power compared to existing alternatives. Unlike tests for marginal genetic association, set-based tests for gene-environment interaction face the challenges of a potentially misspecified and high-dimensional main effect model under the null hypothesis. We show that our proposed test is robust to main effect misspecification of environmental exposure and genetic factors under the gene-environment independence condition. When genetic and environmental factors are dependent, the method of sieves is further proposed to eliminate potential bias due to a misspecified main effect of a continuous environmental exposure. A weighted principal component analysis approach is developed to perform dimension reduction when the number of genetic variants in the set is large relative to the sample size. The methods are motivated by an example from the Multi-Ethnic Study of Atherosclerosis (MESA), investigating interaction between measures of neighborhood environment and genetic regions on longitudinal measures of blood pressure over a study period of about seven years with 4 exams.

CHAPTER I

Introduction

With the advance of high-throughput technologies, high-dimensional genetic data have been widely used in association studies for the identification of genetic variants contributing to common complex diseases. While a large number of genetic variants have been revealed today to be individually associated with complex diseases, they only explain a small proportion of heritability (Manolio, et al., 2009).

Complex diseases are likely influenced by the joint effect of genetic variants through complex biology pathways, given the fact that genes are the functional sets. To improve power and to reduce the burden of multiple comparisons, many genetic association studies have now considered an alternate or supplementary analytic approach towards jointly testing the effect of all SNPs in a biologically defined set, such as a gene, pathway or specific genomic region as opposed to a one-at-a-time single SNP analysis. Aggregation of SNPs is particularly critical for studies of rare variants. A number of methods have gained popularity including kernel machine regression methods (Wu et al. (2011)), similarity regression (Tzeng et al. (2011)) and sum of squared score test (Pan (2009)). Recent studies showed the advantages of these multi-marker tests over individual SNP analyses. First, the genetic markers in LD with the causal SNP(s) carry additional information and may enhance the power of identifying the true effect. Second, gene-based tests considerably reduce the burden of multiple comparisons. Third, Region-based methods are appealing for multi-ethnic cohorts due to differences in LD structure across ethnic groups and thus meta-analysis of a region-based statistic is likely to be more consistent than meta-analysis of single marker tests across ethnicities. Last, gene-based tests enhance the power of identifying rare-variant association in next generation sequencing studies (Morris and Zeggini, 2010).

For genetic studies of cardiovascular disease risk factors, such as the Mulit-Ethnic Study of Atherosclerosis (MESA), observations at multiple time points are available for each individual (Bild, et al., 2002). The longitudinal nature of these studies results in more precise phenotypic characterization, enhancing the ability to associate genes or chromosomal regions with the phenotypes and assess gene-time interaction. However, current statistical methods for testing genetic association in longitudinal studies, in the presence of effect heterogeneity across time are limited, even for one single-nucleotide polymorphism (SNP) at a time analysis (Fan, et al., 2012; Furlotte, Eskin and Eyheramendy, 2012). Investigators often take a simple approach of collapsing the repeated measurements into a single value and hence the method is not able to harness the power of the complete information that is contained in the longitudinal trajectory. One can also apply the standard methods available for correlated outcome models to better utilize the longitudinal data, namely, random effects models (Fitzmaurice, Laird and Ware, 2011) and generalized estimating equations (GEE) (Zeger and Liang, 1986). These methods are primarily proposed for modeling and testing a limited number of SNPs, and cannot be directly applied to assess the joint association of a longitudinally varying outcome with an entire gene or a region with hundreds of SNPs without further modifications.

In addition to genetic association, most complex traits have a multifactorial etiology involving the dynamic interplay of genes and environmental exposures over the life course.

Studies of gene-environment interaction (GEI) often suffer from single one time measurement of exposure or a crude proxy thereof, without proper characterization of lifetime history of cumulative exposure. Longitudinal studies with time varying measures of outcome and exposure data help with characterizing the temporal features of exposure and outcomes, handling exposure measurement error and often enhance power when compared to a cross-sectional analysis. In the context of testing gene-gene/gene-environment interaction for cross-sectional studies, Tzeng et al. (2011), Li et al. (2012), Lin et al. (2013), Chen et al. (2014), Marceau et al. (2015) and Lin et al. (2016) extended the set-based tests for marginal associations to testing interactions. These papers demonstrated superior power of set-based tests for gene-environment interaction by aggregating signals across multiple SNPs. However, no set-based test for gene-environment interaction has been proposed for longitudinal studies where improved power regarding gene-environment interaction is possible by using longitudinally varying outcome and exposure trajectories.

In Chapter II, we propose a random field framework for modeling and testing for the joint association of multiple genetic variants. We view outcomes as stochastic realizations of a random field on a genetic space and propose to use a random field model, referred to as a genetic random field model (GenRF), to model the joint association. This approach is motivated by development in spatial statistics where outcomes are viewed as stochastic realizations of a random field on a Euclidean space (Cressie, 1993). This perspective leads to a very distinctive model from the aforementioned methods; specifically, GenRF regresses the response of one subject on responses of all other subjects. GenRF can be understood from the intuition that genetic similarity leads to trait similarity if variants are associated with the trait. Under the GenRF model, testing for the joint association reduces to a test involving a scalar parameter. Using the pseudo-likelihood method, a test for the joint association is developed, which enjoys many appealing features as SKAT and can achieve comparable or better performance than existing methods. Much of the development is focused on quantitative traits and robustness of the test to other traits, e.g., binary traits, is also discussed.

In Chapter III, we propose a longitudinal genetic random field model (LGRF) and develop generalized score type tests to study the association between repeatedly measured phenotypes and a set of genetic variants in a gene or region. The methods are evaluated through extensive simulation studies and illustrated by analyzing the association between blood pressure and 29 candidate genomic regions across four ethnic groups in MESA.

In Chapter IV, we propose a new statistical approach to test for gene-environment interactions with a set of genetic variants and longitudinally measured outcome and exposure data. The test is robust to misspecification of within subject correlation and is substantially more powerful than an analysis that uses subject-specific averages/summaries of outcome and exposure data. We show that the proposed test is robust to the misspecification of E and G main effects under the gene-environment independence condition. We further propose using the method of sieves to flexibly model the main effect of E for improved type I error control when the gene-environment independence condition does not hold, and better power. We also proposed a weighted principal component analysis (PCA) to remedy the curse of dimensionality when the number of SNPs in the tested set is close or larger than the sample size. We illustrate the proposed methods by both an analysis of targeted GEI (restricted to genetic regions defined around previous GWAS hits) and an agnostic genome-wide gene-based GEI search, with novel time-varying neighborhood features of the environment as exposure, with blood pressure as the longitudinally measured outcome in MESA. Extensive simulation studies, designed to mimic the data structure of MESA are conducted to assess the operating characteristics of the different methods.

CHAPTER II

Set-based Tests for Genetic Association in Cross-sectional Studies

2.1 Introduction

With the advance of high-throughput technologies, high-dimensional genetic data have been widely used in association studies for the identification of genetic variants contributing to common complex diseases. While a large number of genetic variants have been revealed today to be individually associated with complex diseases, they only explain a small proportion of heritability (Manolio et al. (2009)). Complex diseases are likely influenced by the joint effect of genetic variants through complex biology pathways, given the fact that genes are the functional sets. However, the multiple testing problem occurs when one considers a set of single locus analyses, which dramatically diminishes power. Therefore, the joint analysis of a functional set of genetic variants simultaneously can further enhance the discovery process, leading to the identification of new genetic variants associated with complex diseases (Chatterjee et al. (2006)).

Several new statistical methods have been recently developed for joint association analysis, including the kernel machine based method (well known as SKAT)(Wu et al. (2010); Wu et al. (2011)) and the similarity regression (SIMreg) (Tzeng et al. (2009)). Both methods significantly reduce the number of regression parameters, making it feasible and computationally efficient to handle high-dimensional variants. In addition, they account for linkage disequilibrium (LD) and potential interactions, which further improve performance. Both SKAT and SIMreg can be thought of as being developed from the general idea that, if genetic association exists, then genetic similarity leads to trait similarity, which is also the intuition behind our method.

In this project, we propose a random field framework for modeling and testing for the joint association of multiple genetic variants. We view outcomes as stochastic realizations of a random field on a genetic space and propose to use a random field model, referred to as a genetic random field model (GenRF), to model the joint association. This approach is motivated by development in spatial statistics where outcomes are viewed as stochastic realizations of a random field on a Euclidean space (Cressie (2015)). This perspective leads to a very distinctive model from the aforementioned methods; specifically, GenRF regresses the response of one subject on responses of all other subjects. GenRF can be understood from the intuition that genetic similarity leads to trait similarity if variants are associated with the trait. Under the GenRF model, testing for the joint association reduces to a test involving a scalar parameter. Using the pseudo-likelihood method, a test for the joint association is developed, which enjoys many appealing features as SKAT and can achieve comparable or better performance than existing methods, as demonstrated by simulation studies in Section 3 and a real data application in Section 4. Much of the development is focused on quantitative traits and robustness of the test to other traits, e.g., binary traits, is also discussed.

2.2 Method

2.2.1 Genetic Random Field Model

Consider a study where n subjects are sequenced in a region of interest. For subject $i, i = 1, \ldots, n$, let $\boldsymbol{G}_i$ denote the genotype for the p variants within the region, Y_i the trait or phenotype, and $\boldsymbol{X}_i$ the other covariates including, for example, demographic and environmental factors. We are interested in studying the joint association between variants $\boldsymbol{G}_i$ and trait Y_i, possibly adjusted for the effect of $\boldsymbol{X}_i$.

As SKAT and SIMreg, our method is also motivated by the general idea that, if the genetic variants are jointly associated with a trait, then the genetic similarity across subjects will contribute to the trait similarity. To put it in another way, if variants are jointly associated with the trait, then the response of a subject would be close to the response of other subjects who share similar genetic and possibly other variables. Based on this key idea, we propose to directly model the response of each subject as a function of all other responses and the contribution of other responses to Y_i is weighted by their genetic similarity.

For simplicity, we temporarily assume Y_i's are centered (have mean zero) and there are no other adjustment covariates. Specifically, based on the idea discussed above, we model the conditional distribution of Y_i given all other responses as

$$Y_i | \boldsymbol{Y}_{-i} \sim \gamma \sum_{j \neq i} s(\boldsymbol{G}_i, \boldsymbol{G}_j) Y_j + \varepsilon_i, \tag{2.1}$$

where $\boldsymbol{Y}_{-i}$ denotes responses for all other subjects except Y_i ; $s(\boldsymbol{G}_i, \boldsymbol{G}_j)$ is known weights, weighting the contribution of Y_j on approximating (or predicting) Y_i via their genetic similarity; γ is a non-negative coefficient measuring the magnitude of the overall contribution, further discussed below; and ε_i's are random errors. A proper weight function $s(\boldsymbol{G}_i, \boldsymbol{G}_j)$ gives higher value when the two subjects are more similar in terms of genetic variants and, as discussed below, can be viewed as a measure for proximity of two subjects in a genetic space. The random errors ε_i's are assumed to be independent and identically distributed with Normal$(0, \zeta^2)$; extension to distributions other than normal is discussed in Section 2.2.2.

A main distinction between model (2.1) and the usual regression is that (2.1) models the conditional distribution of Y_i given traits of other subjects, whereas in the usual regression one models the conditional distribution of a subject's traits given his/her genetic variants. Intuitively, model (2.1) states that the trait of a subject can be approximated by traits of other subjects who are similar in genetic variants, if variants are associated with the trait. The coefficient γ indicates the magnitude of the trait similarity as a result of genetic similarity. Thus, γ can also be interpreted as a measure for the magnitude of the joint association of $\boldsymbol{G}_i$ with Y_i. Specifically, if $\boldsymbol{G}_i$ is not associated with Y_i, then regardless of how similar subject i is to other subjects in terms of their genetic variants, the trait Y_i is independent of all other Y_j's for $j \neq i$; that is, $\gamma = 0$. On the contrary, if $\boldsymbol{G}_i$ is strongly associated with Y_i, then one may expect Y_i can largely be predicted by traits of subjects having the same or similar genetic variants and a large γ indicates a strong joint association. Therefore, we can test the joint association of genetic variants with the trait by testing a null hypothesis involving a single parameter, i.e., $H_0 : \gamma = 0$.

We have yet to define a measure for "closeness" in the genetic space. Suppose each component of $\boldsymbol{G}_i$ records the number of minor alleles in a single locus and takes on values $\{0, 1, 2\}$, respectively, corresponding to $\{AA, Aa, aa\}$. Then a sensible measure for closeness or similarity is the so called identity-by-state (IBS) (Wu et al. (2010)), defined as

$$s(\boldsymbol{G}_i, \boldsymbol{G}_j) = \sum_{k=1}^{p} \{2 - |G_{ik} - G_{jk}|\}.$$

That is, the IBS measures the number of alleles in the region of interest shared by two individuals; for example, for $p = 1$, $s(AA, AA) = 2, s(Aa, aa) = 1, s(AA, aa) = 0$. Other measures for closeness in the genetic space rather than IBS are also possible, e.g., the other kernel functions discussed in Wu et al. (2010), providing flexibility in our GenRF model. Similar to SKAT, our GenRF model can also incorporate weights to increase the importance of rare variants. Specifically, one can define $s(\boldsymbol{G}_i, \boldsymbol{G}_j) = \sum_{k=1}^{p} w_k \{2 - |G_{ik} - G_{jk}|\}$, where w_k is a prespecified weight for variant k; see Wu et al. (2010) for more discussions on w_k.

So far we have focused on the situation where no covariate adjustment is required. If adjustment for other factors is needed a natural extension of model (2.1) is given by

$$Y_i|\boldsymbol{Y}_{-i}, \boldsymbol{X}_i \sim \boldsymbol{\beta}^T\boldsymbol{X}_i + \gamma\sum_{j\neq i} s(\boldsymbol{G}_i, \boldsymbol{G}_j)(Y_j - \boldsymbol{\beta}^T\boldsymbol{X}_j) + \varepsilon_i. \tag{2.2}$$

An intercept term is included in $\boldsymbol{X}_i$ and, as a result, in (2.2) Y_i's are not required to be centered. Under this model, testing for the joint association of $\boldsymbol{G}_i$ with Y_i after adjusting for other factors is also equivalent to testing $H_0 : \gamma = 0$. We will mainly focus on this more general form of the GenRF model in the development of a testing procedure. For simplicity, the matrix form of the GenRF model is given by

$$\boldsymbol{Y}|\boldsymbol{Y}_{-}, \boldsymbol{X} = \boldsymbol{X\beta} + \gamma\boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{X\beta}) + \boldsymbol{\varepsilon}, \tag{2.3}$$

where $\boldsymbol{Y}$ is $(Y_1, \ldots, Y_n)^T$; $\boldsymbol{Y}_{-}$ is $(\boldsymbol{Y}_{-1}, \ldots, \boldsymbol{Y}_{-n})^T$; $\boldsymbol{X}$ is an $n \times q$ matrix defined as $(\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$; $\boldsymbol{\varepsilon} \sim$ Normal $(0, \zeta^2 I_{n\times n})$; and $\boldsymbol{S}$ is an $n \times n$ symmetric matrix with zeros on the diagonal and the (i, j)-th element $s(\boldsymbol{G}_i, \boldsymbol{G}_j)$ for $i \neq j$.

According to the factorization theorem of Besag (1974), our GenRF model in (2.2) uniquely determines the following joint distribution of $\boldsymbol{Y}$, i.e.,

$$\boldsymbol{Y}|\boldsymbol{X} \sim \boldsymbol{X\beta} + \boldsymbol{v}, \;\; \boldsymbol{v} \sim N(0, \zeta^2(\boldsymbol{I} - \gamma\boldsymbol{S})^{-1}), \tag{2.4}$$

where $\boldsymbol{v}$ is an n-dimensional random column vector. Note, the coefficient γ used for describing the conditional expectation of Y_i given others in model (2.1) actually describes the correlations among Y_i's. It is clear that, under the null hypothesis that there is no association between $\boldsymbol{G}_i$ and Y_i, i.e., $\gamma = 0$, Y_i's are uncorrelated, but if $\gamma > 0$, GenRF states that Y_i's are positively correlated as a result of having similar genetic variants associated with the trait.

2.2.2 Genetic Random Field Test

In this subsection, we focus on developing a test for the null hypothesis $H_0 : \gamma = 0$ based on model (2.2). Model (2.2) states that, given responses from all other subjects and covariates $\boldsymbol{X}_i$, the conditional distribution of Y_i is normal with mean $\boldsymbol{\beta}^T\boldsymbol{X}_i + \gamma\sum_{j\neq i} s(\boldsymbol{G}_i, \boldsymbol{G}_j)(Y_j - \boldsymbol{\beta}^T\boldsymbol{X}_j^T)$ and variance ζ^2. We construct the pseudo-likelihood according to Besag (1975) as

$$L_{pd} = \prod_{i=1}^{n}\Big\{\frac{1}{\sqrt{2\pi\zeta^2}}\exp\Big[-\frac{1}{2\zeta^2}\big\{Y_i - \boldsymbol{\beta}^T\boldsymbol{X}_i - \gamma\sum_{j\neq i} s(\boldsymbol{G}_i, \boldsymbol{G}_j)(Y_j - \boldsymbol{\beta}^T\boldsymbol{X}_j)\big\}^2\Big]\Big\},$$

which is a product of the conditional densities of Y_i across i. Also according to Besag (1975), assuming $\boldsymbol{\beta}$ is known, one may estimate γ by the maximum pseudo-likelihood method. The estimator for γ can be obtained by minimizing

$$\sum_{i=1}^{n}\big\{Y_i - \boldsymbol{\beta}^T\boldsymbol{X}_i - \gamma\sum_{j\neq i} s(\boldsymbol{G}_i, \boldsymbol{G}_j)(Y_j - \boldsymbol{\beta}^T\boldsymbol{X}_j\big\}^2,$$

which in matrix notation is equal to

$$\{(\boldsymbol{I} - \gamma\boldsymbol{S})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\}^T(\boldsymbol{I} - \gamma\boldsymbol{S})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The minimization leads to an estimator for γ given by

$$\Rightarrow \widetilde{\gamma} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{S}^2(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}. \tag{2.5}$$

Intuitively one expects that a large value of $\widetilde{\gamma}$ would give us evidence to reject the null hypothesis that $\gamma = 0$. In practice, $\boldsymbol{\beta}$ in unknown. We propose to replace $\boldsymbol{\beta}$ by its least square estimator $\widehat{\boldsymbol{\beta}}$ under the null hypothesis, i.e., $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ which is unbiased for $\boldsymbol{\beta}$. Substitute $\widehat{\boldsymbol{\beta}}$ into the expression for $\widetilde{\gamma}$ and straightforward algebra leads to the final test statistic:

$$\widehat{\gamma} = \frac{\boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{S}\boldsymbol{B}\boldsymbol{Y}}{\boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{S}^2\boldsymbol{B}\boldsymbol{Y}}, \tag{2.6}$$

where $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Again a large value of $\widehat{\gamma}$ supports the rejection of the null hypothesis.

We next show how the p-value for testing $\gamma = 0$ can be obtained based on the test statistic $\widehat{\gamma}$; i.e., we would like to calculate the probability of $\widehat{\gamma}$ greater than the observed value of the statistic under the null hypothesis. Suppose η is the observed value of the test statistic $\widehat{\gamma}$. Since $\boldsymbol{B}\boldsymbol{S}^2\boldsymbol{B}$ is positive-definite, we have

$$P_{H_0}\left(\frac{\boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{S}\boldsymbol{B}\boldsymbol{Y}}{\boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{S}^2\boldsymbol{B}\boldsymbol{Y}} > \eta\right) = P_{H_0}\Big((\boldsymbol{B}\boldsymbol{Y})^T(\boldsymbol{S} - \eta\boldsymbol{S}^2)\boldsymbol{B}\boldsymbol{Y} > 0\Big)$$

As it is assumed that $\varepsilon_i \sim N(0, \zeta^2)$, i.i.d. across i, it follows that $\boldsymbol{B}\boldsymbol{Y} \sim N(0, \zeta^2\boldsymbol{B}^2)$ under the null hypothesis. On the other hand, the statistic $\widehat{\gamma}$ in (2.6) is ancillary to ζ^2 because ζ^2 in the numerator and denominator cancels out. Therefore, the above equation becomes

$$P_{H_0}\Big((\boldsymbol{B}\boldsymbol{Y})^T(\boldsymbol{S} - \eta\boldsymbol{S}^2)\boldsymbol{B}\boldsymbol{Y} > 0\Big) = P\Big(\boldsymbol{Z}^T(\boldsymbol{S} - \eta\boldsymbol{S}^2)\boldsymbol{Z} > 0\Big),$$

where $\boldsymbol{Z}$ is an $n \times 1$ random vector following $N(0, \boldsymbol{B}^2)$. Applying standard results on the distribution of quadratic form in normal random variables, we have

$$\boldsymbol{Z}^T(\boldsymbol{S} - \eta\boldsymbol{S}^2)\boldsymbol{Z} \sim \sum_i^n \lambda_i \Phi_i,$$

where Φ_i's are i.i.d random variables with χ^2_1 distribution, and $\{\lambda_i\}$ are the eigenvalues of $\boldsymbol{B}(\boldsymbol{S} - \eta\boldsymbol{S}^2)\boldsymbol{B}$. The final p-value can be obtained by Davies' exact method (Davies (1980)) for the weighted summation of independent Chi-square variables.

The proposed test has several appealing properties. First, due to the analytical form of the test statistic, the computational burden is well controlled. Second, as $\widehat{\gamma}$ in (2.6) is ancillary to ζ^2, unlike SKAT, there is no need to plug in a consistent estimator for ζ^2. Third, the proposed method improves power by exploiting LD and allowing for possible complex interactions among variants. Linkage disequilibrium can cause correlations between variants, especially when we consider nearby loci. Considering similarity in variants can naturally reduce the degree of freedom. In the extreme case where components of $\boldsymbol{G}_i$ are "perfectly correlated", the similarity argument will consider the whole set as a single variable. Finally, as SKAT, the GenRF test can boost power of testing rare variants by increasing their weights by specifying w_k appropriately for variant k.

2.2.3 Robustness to Other Distributions

The derivation of the GenRF test given above is built on the normal distribution assumption. Asymptotically, the proposed test is robust to distributions other than normal with slight modification. Consider $P_{H_0}\Big((\boldsymbol{BY})^T(\boldsymbol{S}-\eta\boldsymbol{S}^2)\boldsymbol{BY} > 0\Big)$, where it is now assumed $\boldsymbol{Y}$ follows an arbitrary distribution with mean zero and possibly heteroscadastic variances. The random quantity $(\boldsymbol{BY})^T(\boldsymbol{S}-\eta\boldsymbol{S}^2)\boldsymbol{BY}$ is a quadratic form in $\boldsymbol{BY}$ (with mean 0) with matrix $\boldsymbol{A} = \boldsymbol{S}-\eta\boldsymbol{S}^2$. Rotar (1974) proved that under sufficiently weak conditions on matrix $\boldsymbol{A}$ and for large n, $P_{H_0}\big((\boldsymbol{BY})^T(\boldsymbol{S}-\eta\boldsymbol{S}^2)\boldsymbol{BY} > 0\big)$ is close to $P_{H_0}(\widetilde{\boldsymbol{Z}}^T(\boldsymbol{S}-\eta\boldsymbol{S}^2)\widetilde{\boldsymbol{Z}} > 0)$, where $\widetilde{\boldsymbol{Z}}$ follows $N(0,\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being the covariance matrix of $\boldsymbol{BY}$. In addition, Gotze and Tikhomirov (1999) gave an upper bound on $\sup_x \big|P_{H_0}\big((\boldsymbol{BY})^T\boldsymbol{ABY} < x\big) - P_{H_0}(\widetilde{\boldsymbol{Z}}^T\boldsymbol{A}\widetilde{\boldsymbol{Z}} < x)\big|$. These properties lead to the robustness of the GenRF test, with minor modification, as long as $\boldsymbol{BY}$ has expectation zero under the null hypothesis, which is true since the least squares estimator $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ is unbiased for the mean of $\boldsymbol{Y}$ regardless of the distribution of $\boldsymbol{Y}$. For example, for binary traits, $\boldsymbol{\Sigma} = \boldsymbol{BWB}$, where $\boldsymbol{W} = \text{diag}(\mu_1(1-\mu_1),\ldots,\mu_n(1-\mu_n))$ and $\mu_i = \boldsymbol{\beta}^T\boldsymbol{X}_i$. Then

$$\widetilde{\boldsymbol{Z}}^T(\boldsymbol{S}-\eta\boldsymbol{S}^2)\widetilde{\boldsymbol{Z}} \sim \sum_i^n \widetilde{\lambda}_i\Phi_i,$$

where Φ_i's are i.i.d random variables with χ_1^2 distribution, and $\{\widetilde{\lambda}_i\}$ are the eigenvalues of $\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{B}(\boldsymbol{S}-\eta\boldsymbol{S}^2)\boldsymbol{BW}^{\frac{1}{2}}$. The final p-value can be also obtained by Davies' exact method (Davies (1980)). We comment that, as the score test in SKAT is of similar quadratic form, one would expect that SKAT may share this property as well.

Therefore, one can directly use the test statistic in (2.6) for binary traits or traits that have distributions other than normal and the test, with a minor modification on the null distribution considering heteroscedastic variances, would be asymptotically valid. Note, this test corresponds to a model where the trait mean is related to a linear predictor through an identity link and may seem unnatural for binary traits. However, we argue that the model is mostly viewed as a mean leading to a sensible test. We also note that the commonly used trend test for testing genetic associations in an additive genetic model can be developed from a linear model for the mean of a binary trait (Laird and Lange (2010)), and a linear model is used for testing genetic associations for a binary trait in Ballard et al. (2010) as well. We note that a possible practical issue for binary traits may arise in practice, i.e., the estimated means $\{\widehat{\mu}_i\}$ may be outside of [0,1] and consequently $\widehat{\boldsymbol{W}}^{\frac{1}{2}}$is not well defined. In this case, a remedy is to truncate the predictions $\{\widehat{\mu}_i\}$ at 0 or 1. The practical issue may arise when covariates have a wide support and a very strong effect and is less of a concern otherwise, for example, when covariates are categorical. Certainly, studying other link functions, e.g., the logit link, to avoid this practical problem is important in the future. The validity of the test, corresponding to an identity link, is further studied by simulations shown in sections 1 and 2 of the Supplementary Materials (Appendix A).

2.3 Application: Dallas Heart Study

We applied our method to the Dallas Heart Study (Browning et al. (2004)), a population-based, multi-ethnic study on 3551 subjects whose Lipids and glucose metabolism were measured. In this study, 348 sequence variations in the coding regions of the four genes, ANGPTL3, ANGPTL4, ANGPTL5 and ANGPTL6 were discovered. Most of these variants (86%) are rare with MAF less than 1%. More information regarding the number of rare variants is shown in the Supplementary Materials (Appendix A).

Individuals who have diabetes mellitus, alcohol dependency or have taken lipids lowering drugs were excluded as these factors may confound the interpretation of associations. Our final analysis was based on data on 2812 subjects after quality control steps.

We assessed the association between ANGPTL gene families and two traits, specifically high-density lipoprotein (HDL) and triglyceride, using the proposed GenRF test and SKAT, both with and without weighting. As in the simulation studies, the IBS kernel and the Beta (1, 25) weight were applied. Analyses were also carried out using the more traditional methods including PCR, MinSNP, VT and F-test. Our analysis were done for the non-synonymous variants, adjusted for gender and ethnicity.

The association between ANGPTL4 gene and the level of HDL and triglyceride was previously discovered by Romeo et al. (2007). In our analysis, both weighted GenRF and SKAT gave evidence for the ANGPTL4 and triglyceride association (p-values: 0.019 and 0.006). Among all the methods considered, only weighted SKAT showed marginal evidence for the association between ANGPTL4 and HDL (p-value: 0.040). One possible explanation is that the causal proportion of ANGPTL4 is low and SKAT performs better in this case as shown in simulation studies. Note that the weighted GenRF and SKAT uncovered these associations while the unweighted tests did not, possibly indicating the causal variants in ANGPTL4 might be rare (MAF $< 5\%$), or the effect size is negatively correlated with allele frequency. As for ANGPTL5, our analysis using GenRF provided evidence to support the association with HDL (p-value: 0.009 and 0.036 for weighted and unweighted analyses) while SKAT provided marginal evidence (p-value: 0.035 and 0.050). Note the unweighted tests gave larger p-values. Since all variants in ANGPTL5 are rare (MAF $< 5\%$), the result suggests that the causal variants might be the rare variants with relatively higher allele frequency. This finding was supported by standard approaches

Table 2.1: Application to Dallas Heart Study for non-synonymous variants. GenRF: the unweighted genetic random field test; SKAT: the unweighted sequential kernel association test of Wu, et al. (2011); PCR: the princeple component regression test of Guaderman et al. (2007); MinSNP: the MinSNP test considered by Ballard et al. (2010); F-test: the F-test in linear regression; GenRF.w: the genetic random field test with Beta (1, 25) weight of Wu, et al. (2011); SKAT.w: the sequential kernel association test with Beta (1, 25) weight; VT: the variable-threshold test of Price et al. (2010). $*$ indicates p-value is less than or equal to $\alpha = 0.05$.

Method	P-value			
	HDL			
	ANGPTL3	ANGPTL4	ANGPTL5	ANGPTL6
GenRF	0.487	0.181	0.009∗	0.417
SKAT	0.981	0.423	0.035∗	0.504
PCR	0.980	0.775	0.197	0.434
MinSNP	0.178	0.329	0.033∗	0.729
F-test	0.331	0.148	0.051	0.786
GenRF.w	0.345	0.218	0.036∗	0.496
SKAT.w	0.965	0.040∗	0.050∗	0.535
VT	0.393	0.111	0.051	0.488
	Triglyceride			
	ANGPTL3	ANGPTL4	ANGPTL5	ANGPTL6
GenRF	0.025∗	0.221	0.428	0.857
SKAT	0.050∗	0.312	0.936	0.755
PCR	0.129	0.780	0.787	0.762
MinSNP	0.562	0.219	0.921	0.713
F-test	0.587	0.380	0.904	0.530
GenRF.w	0.100	0.019∗	0.180	0.466
SKAT.w	0.075	0.006∗	0.906	0.756
VT	0.993	0.905	0.968	0.050∗

like MinSNP (p-value: 0.033), F-test (p-value: 0.051) and VT test (p-value: 0.051). More results are shown in table 4. Overall, for this study, GenRF performs comparably to SKAT and seems to perform better than the other more standard methods.

2.4 Simulation Studies

We report results of several simulations, each based on 1000 Monte Carlo (MC) replicates, to evaluate the performance of the GenRF test, relative to existing methods including SKAT. Four sets of simulations are conducted to evaluate 1) type-1 error rates under different minor allele frequencies (MAF) and sample sizes, 2) power for common variant analysis under different LD, interaction effect, and proportions of causal SNPs, 3) power under scenarios where the causal SNPs include rare variants, and 4) robustness of the GenRF test to different distributions of the response variable.

In the first set of simulations, we evaluated type-I error rates using sample size $n = 50$, 100, 200 and 500 . Genotypes for $p = 20$ loci without LD were simulated, with MAF for each locus 0.005, 0.01, 0.1, or 0.2. Responses were generated according to

$$Y_i = \varepsilon_i, \text{ where } \quad \varepsilon_i \sim N(0,1),$$

so that no genetic variant is associated with the trait.

In the second set of simulations, we evaluated power under scenarios varying in LD, interaction effects, or the proportions of causal SNPs, setting $p = 20$ and $n = 500$. To simulate LD, the 20 loci were evenly divided into two regions. For each region, the haplotype allele was simulated one by one with MAF 0.2 and correlation coefficient (ρ) between adjacent pair of alleles equal to 0, 0.2, 0.4, 0.8 respectively for each scenario. Genotypes were then generated by summing up two haplotype vectors. This way, all the loci are positively correlated with others in the same region. Responses were generated according to

$$Y_i = 0.2G_{i,5} + 0.2G_{i,15} + \varepsilon_i, \text{ where } \quad \varepsilon_i \sim N(0,1).$$

That is, variants 5 and 15, belonging to different LD regions, are associated with the trait.

To generate data with complex interactions, we set MAF 0.2, and the LD parameter $\rho = 0.4$. Data were generated such that two-way interactions exist between K ($K = 1, 2, 3$ or 4) pairs of alleles, with alleles in each pair belonging to the two different LD regions as described above. Responses were then generated according to the following model,

$$Y_i = 0.2\sum_{k=1}^{K} G_{i,4+k}G_{i,14+k} + \varepsilon_i, \text{ where } \quad \varepsilon_i \sim N(0,1). \tag{2.7}$$

We see that these models contain only interactions but no main effect of each locus.

To examine the effect of causal proportion, we set MAF 0.2 and $\rho = 0.4$. For each MC data set, K causal SNPs were randomly selected with $K = 1, 2, 3$, or 4, each corresponding to 5%, 10%, 15% and 20% causal SNPs. Responses were then generated according to

$$Y_i = 0.15 \sum_{k=1}^{K} G_{i,B_k} + \varepsilon_i, \text{ where } \quad \varepsilon_i \sim N(0,1), \tag{2.8}$$

where $(G_{B_1}, \ldots, G_{B_K})$ are the selected causal SNPs.

Simulation set 2 has focused on common variants. The third set of simulations considered scenarios involving rare variants and the scenarios vary in proportions of causal variants. We set $p = 20$, $n = 500$, and $\rho = 0$. The 20 SNPs were divided into two regions, one with 16 rare variants (MAF 0.008) and one with 4 common variants (MAF 0.1). Note, the proportion of rare variants is chosen according to the Dallas Heart Study . Two scenarios were considered where traits were associated with: 1) rare variants only or 2) both common and rare variants. For each scenario, K rare SNPs were causal with $K = 1$, 2, 4, 6, 8, 10, 12 or 14, i.e., $K \times 6.25\%$ SNPs in the rare region are causal. In the scenario that both rare and common variants are causal, we set one of the common SNP as causal additionally. The effect size β was set to be a decreasing function of MAF with $\beta = 0.2 \times |\log_{10} \text{MAF}|$ as in Wu et al. (2011). Responses were generated according to the following model,

$$Y_i = \beta_1 \sum_{k=1}^{K} G_{i,k} + \beta_2 G_{i,20} + \varepsilon_i, \text{ where } \quad \varepsilon_i \sim N(0,1),$$

where $\beta_1 = 0.2 \times |\log_{10} 0.008|, \beta_2 = 0$ for scenario 1 and $\beta_1 = 0.2 \times |\log_{10} 0.008|, \beta_2 = 0.2 \times |\log_{10} 0.1|$ for scenario 2.

We considered one additional scenario where the 500 subjects' genotypes were simulated based on data from the Dallas Heart Study. For each MC data set, we randomly selected one gene, then we randomly choose $10\%, 20\%, \ldots, 80\%$ causal variants from those rare variants with true MAF less than 1%. Traits were simulated by

$$Y_i = \sum_{k=1}^{K} \beta_{B_k} G_{i,B_k} + \varepsilon_i, \text{ where } \quad \varepsilon_i \sim N(0,1),$$

where $(G_{B_1}, \dots, G_{B_K})$ are the selected causal variants and $\beta_{B_k} = 0.2 \times |\log_{10} \text{MAF}_{B_k}|$.

In the fourth set of simulations, we further evaluated the robustness of the GenRF test to distributions other than normal, specifically, exponential, binary and mixture normal distributions. Details on the simulation setup is described in the Supplementary Materials (Appendix A).

In terms of type-I error rates, we only evaluated the proposed GenRF test and SKAT. In both GenRF and SKAT, we adopted the IBS kernel and considered both weighted and unweighted (i.e., $w_k = 1$) versions; in the weighted version, Beta (1, 25) weight as in Wu et al. (2011) was used. In addition to SKAT, we compared GenRF test to other more standard methods. For common variant scenarios, we included the principle component regression test (PCR) (Gauderman et al. (2007)); the MinSNP test (Ballard et al. (2010)) and the F-test in linear regression model including only main effects. For scenarios involving rare variants, the variable-threshold (VT) test (Price et al. (2010)) was included.

Table 1 shows results for the first set of simulations with different MAF and sample sizes. The GenRF test achieves the type I error rate close to the nominal level. However, SKAT is conservative in some scenarios due to the estimation of nuisance parameters, especially when the sample size is small. Since the GenRF test is an exact test without asymptotic approximation under normal assumption, the type I error rate is better controlled.

Table 2 shows the power of various methods under common variant scenarios. The first part shows the effect of LD on power. When LD does not exist or is low, e.g., $\rho < 0.4$, the three linear regression based tests, PCR, MinSNP and F-test , are more powerful as expected because the data were generated exactly from a linear model. Among them, the PCR and MinSNP can exploit LD and have increasing power when LD is higher. When LD is moderate or high, both the GenRF test and SKAT have higher or even substantially higher power than the other tests by borrowing information from other loci. The power of the GenRF test is comparable to that of SKAT.

The second part shows results when there are complex interactions between variants but no main effects. Note the LD structure is the same as that in part 1 with $\rho = 0.4$ in which the five methods have comparable power. Therefore, the power difference is mainly due to the complex interactions. In these scenarios, the linear regression based methods has low power in detecting the joint association. Both GenRF test and SKAT attain much larger power. Moreover, the proposed GenRF test has larger power than SKAT in detecting the joint association effect when complex interactions exist.

The third part shows results when the causal proportion varies. Similarly, the LD parameter ρ is set to be 0.4 to eliminate the impact of factors other than the causal proportion. Because MinSNP is based on single SNP analysis, the test is less powerful especially when causal proportion is high, i.e. 15% or 20%. GenRF and SKAT show comparable power in general, but GenRF performs better as causal proportion gets higher.

Table 3 shows results for scenarios involving rare variants . When the trait is only associated with rare variants, the weighted GenRF and SKAT have significantly larger power as we expected because the weights favor the rare variants. The weighted GenRF has lower power than SKAT when the causal proportion is low, e.g., $\leq 25\%$, but has larger power then the proportion is greater than 25%. Both weighted GenRF and SKAT have comparable or larger power relative to the VT test and F-test. The scenario based on the Dallas Heart Study shows similar results, i.e. GenRF performs better under higher causal proportion ($\geq 20\%$).

When causal variants include both common and rare variants and the effect size is a decreasing function of MAF, the unweighted GenRF and SKAT have comparably larger power than the weighted tests when the rare causal proportion is low ($\leq 37.5\%$). This is not surprising as the effect of the common variant is relatively large but down-weighted in the weighted GenRF and SKAT. As the rare causal proportion increases and the number of common variants is fixed at one, the results change dramatically. When the rare causal proportion is higher than 37.5%, the weighted GenRF and SKAT show higher power than the unweighted counterpart. Overall, for scenarios considered here, the GenRF test has very good performance relative to others.

Table 2.2: Type I error rate simulation results under different levels of MAF and sample size (1000 replicates). Each cell contains the type I error rate, i.e., rejection rate when data are generated under the null model. GenRF: the unweighted genetic random field test; SKAT: the unweighted sequential kernel association test of Wu, et al. (2011); GenRF.w: GenRF with Beta(1,25) weight as in Wu, et al. (2011); SKAT.w: SKAT with Beta(1,25) weight.

Methods		Different Levels of MAF and Sample Size (n)							
	MAF	0.005				0.01			
	n	50	100	200	500	50	100	200	500
GenRF		0.043	0.049	0.050	0.045	0.040	0.043	0.061	0.048
GenRF.w		0.048	0.056	0.051	0.045	0.043	0.046	0.060	0.046
SKAT		0.035	0.051	0.057	0.057	0.034	0.046	0.050	0.039
SKAT.w		0.034	0.050	0.053	0.059	0.029	0.042	0.046	0.035
	MAF	0.1				0.2			
	n	50	100	200	500	50	100	200	500
GenRF		0.051	0.049	0.055	0.044	0.050	0.058	0.055	0.049
GenRF.w		0.052	0.052	0.053	0.048	0.047	0.051	0.052	0.046
SKAT		0.022	0.039	0.041	0.041	0.016	0.030	0.041	0.043
SKAT.w		0.041	0.035	0.043	0.054	0.045	0.043	0.046	0.041

Table 2.3: Power simulation results for common variant analysis under different levels of linkage disequilibrium (LD), interaction effects and causal proportion (1000 replicates). GenRF: the unweighted genetic random field test; SKAT: the unweighted sequential kernel association test of Wu, et al. (2011); PCR: the princeple component regression test of Guaderman et al. (2007); MinSNP: the MinSNP test considered by Ballard et al. (2010); F-test: the F-test in linear regression.

Method	Different Level of LD				Number of Two-way Interactions				Different Causal Proportion			
	0	0.2	0.4	0.8	1	2	3	4	5%	10%	15%	20%
GenRF	0.462	0.472	0.566	0.816	0.119	0.364	0.652	0.862	0.124	0.321	0.539	0.776
SKAT	0.491	0.487	0.545	0.764	0.100	0.299	0.546	0.746	0.150	0.324	0.506	0.727
PCR	0.495	0.467	0.518	0.676	0.119	0.268	0.470	0.657	0.159	0.308	0.473	0.679
MinSNP	0.570	0.507	0.543	0.656	0.098	0.252	0.408	0.576	0.180	0.342	0.463	0.624
F-test	0.545	0.514	0.524	0.538	0.112	0.231	0.394	0.562	0.145	0.278	0.471	0.665

Table 2.4: Power simulation results under scenarios involving rare variants with different proportion of causal variants (1000 replicates). Rare Causal Variants: causal variants are rare only; Common & Rare Causal Variants: causal variants are both rare and common; DHS: scenario based on the Dallas Heart Study. GenRF.w: the genetic random field test with Beta (1, 25) weight as in Wu, et al. (2011); SKAT.w: the sequential kernel association test with Beta (1, 25) weight; VT: the variable-threshold test of Price et al. (2010); Other entries as in Table 2.

Method	Different Proportion of Causal Variants							
	Rare Causal Variants							
	6.25%	12.5%	25%	37.5%	50%	62.5%	75%	87.5%
GenRF	0.045	0.073	0.139	0.209	0.305	0.466	0.593	0.739
SKAT	0.048	0.052	0.066	0.073	0.086	0.100	0.111	0.126
GenRF.w	0.062	0.087	0.212	0.429	0.660	0.848	0.950	0.980
SKAT.w	0.083	0.125	0.252	0.368	0.515	0.654	0.736	0.814
VT	0.065	0.082	0.128	0.209	0.314	0.487	0.680	0.852
F-test	0.080	0.113	0.190	0.302	0.449	0.556	0.670	0.765
	Common & Rare Causal Variants							
	6.25%	12.5%	25%	37.5%	50%	62.5%	75%	87.5%
GenRF	0.191	0.259	0.380	0.501	0.625	0.761	0.861	0.927
SKAT	0.274	0.281	0.287	0.313	0.331	0.359	0.387	0.416
GenRF.w	0.061	0.097	0.232	0.434	0.646	0.853	0.939	0.981
SKAT.w	0.078	0.155	0.277	0.386	0.523	0.631	0.732	0.818
VT	0.217	0.306	0.418	0.504	0.603	0.720	0.845	0.930
F-test	0.163	0.270	0.354	0.477	0.618	0.701	0.779	0.843
	DHS							
	10%	20%	30%	40%	50%	60%	70%	80%
GenRF	0.080	0.140	0.169	0.247	0.329	0.414	0.507	0.600
SKAT	0.071	0.089	0.114	0.117	0.153	0.191	0.205	0.271
GenRF.w	0.100	0.204	0.321	0.434	0.588	0.696	0.796	0.875
SKAT.w	0.118	0.196	0.294	0.330	0.433	0.544	0.600	0.688
VT	0.095	0.159	0.254	0.359	0.498	0.612	0.721	0.827
F-test	0.147	0.239	0.355	0.423	0.528	0.653	0.721	0.795

CHAPTER III

Set-based Tests for Genetic Association in Longitudinal Studies

3.1 Introduction

Genome-wide association studies (GWAS) have been successful in identifying susceptibility loci for risk factors of chronic diseases. For genetic studies of cardiovascular disease risk factors, such as the Mulit-Ethnic Study of Atherosclerosis (MESA), observations at multiple time points are available for each individual (Bild et al. (2002)). The longitudinal nature of these studies results in more precise phenotypic characterization, enhancing the ability to associate genes or chromosomal regions with the phenotypes and assess gene-time interaction. However, current statistical methods for testing genetic association in longitudinal studies, in the presence of effect heterogeneity across time are limited, even for one single-nucleotide polymorphism (SNP) at a time analysis (Fan et al. (2012); Furlotte et al. (2012)). Investigators often take a simple approach of collapsing the repeated measurements into a single value and hence the method is not able to harness the power of the complete information that is contained in the longitudinal trajectory. One can also apply the standard methods available for correlated outcome models to better utilize the longitudinal data, namely, random effects models (Fitzmaurice et al. (2012)) and generalized estimating equations (GEE) (Liang and Zeger (1986)). These methods are primarily proposed for modeling and testing a limited number of SNPs, and cannot be directly applied to assess the joint association of a longitudinally varying outcome with an entire gene or a region with hundreds of SNPs without further modifications.

Recent studies showed the advantages of multi-marker tests over individual SNP analyses. First, the genetic markers in LD with the causal SNP(s) carry additional information and may enhance the power of identifying the true effect. Second, gene-based tests considerably reduce the burden of multiple comparisons. Third, Region-based methods are appealing for multi-ethnic cohorts due to differences in LD structure across ethnic groups and thus meta-analysis of a region-based statistic is likely to be more consistent than meta-analysis of single marker tests across ethnicities. Last, gene-based tests enhance the power of identifying rare-variant association in next generation sequencing studies (Morris and Zeggini (2010)). Two notable existing approaches are the sequence kernel association tests (SKAT) (Wu et al. (2011)) and similarity regression (SIMreg) (Tzeng et al. (2009)). From a random field framework and borrowing ideas from spatial statistics, the genetic random field model (GenRF) was recently developed for modeling and testing joint associations (He et al. (2014); Li et al. (2014)). So far, however, extensions of these methods are not available for longitudinal data.

It is desirable to have a multi-marker test for longitudinal studies that can incorporate the time-dependent variation in outcome, utilize all the variants in a gene or region and boost power in the presence of effect heterogeneity across time. Extending the GenRF method to the longitudinal setting, we propose a longitudinal genetic random field model (LGRF) and develop generalized score type tests to study the association between repeatedly measured phenotypes and a set of genetic variants in a gene or region. The methods are evaluated through extensive simulation studies and illustrated by analyzing the association between blood pressure and 29 candidate genomic regions across four ethnic groups in MESA.

3.2 Method

Consider a study population of m subjects, and the i-th subject has n_i repeated observations. Each subject is sequenced in a region of interest with q variants, and measured on p additional non-genetic covariates such as age, gender and other potential confounders. Let $Y_{i,l}$ be the phenotypic value for the l-th observation on the i-th subject, measured at time $t_{i,l}$; $\boldsymbol{G}_i = (G_{i,1}, G_{i,2}, \ldots, G_{i,q})^T$ be the genotypes for the q variants within the region where $G_{i,h} \in \{0, 1, 2\}$ for any $1 \leq h \leq q$, which does not change over time; $\boldsymbol{X}_{i,l} = (X_{i,l,1}, \ldots, X_{i,l,p})^T$ be the covariates corresponding to the l-th observation on the i-th subject, either time-varying or time-invariant. We denote $n = \sum_i n_i$, $\boldsymbol{Y}_{n\times 1} = (Y_{1,1}, \ldots, Y_{1,n_1}, Y_{2,1}, \ldots)^T$ and define $\boldsymbol{X}_{n\times p}$, $\boldsymbol{G}_{n\times q}$ similarly for covariates and genotypes. We are interested in investigating the association between phenotype $\boldsymbol{Y}_{n\times 1}$ and variants $\boldsymbol{G}_{n\times q}$, adjusted for the effect of $\boldsymbol{X}_{n\times p}$.

3.2.1 Longitudinal Genetic Random Field Model

The GenRF method (He et al. (2014)) is a gene-based association test motivated by the general idea that, if the genetic variants in a region are jointly associated with the phenotype, then subjects having similar genotypes in that region will have similar phenotype (Tzeng et al. (2009)). Motivated by development in spatial statistics (Cressie (2015)) and random field theory (Besag (1974); Adler and Taylor (2009)), GenRF views phenotypic values as a random field on a genetic space where the vector of genotype sequences determines the location in the space; i.e., the phenotype at each location is a random variable and these random variables are possibly correlated depending on their spatial location, e.g., the closer the more similar. It directly regresses the phenotype of a given subject on that of all others, where the contribution of other subjects is weighted by their genotype similarity with the given subject. This leads to a conditional autoregressive model commonly used in spatial statistics to study spatial dependence.

With repeated measurements, one has to appropriately account for the within-subject correlation between outcomes to obtain valid inference and improve efficiency. Extending the GenRF model to the longitudinal setting, we propose a longitudinal GenRF (LGRF) model, where the conditional mean of each observation is modeled as a weighted sum of all other observations, including those from the same subject. In a longitudinal setting, one may expect that phenotypes from the same subject may be more similar due to reasons other than the shared genetic variants of interest. To capture this, we define a within-subject similarity, which depends on the time between two measurements on the same subject; for example, if two observations are measured closer in time, their within-subject similarity may be larger. Formally LGRF model is written as:

(3.1)

$$Y_{i,l}|\boldsymbol{Y}_{-(i,l)} = \boldsymbol{X}_{i,l}^T\boldsymbol{\beta}+\sum_{k\neq l} w(t_{i,k},t_{i,l};\boldsymbol{\eta})(Y_{i,k}-\boldsymbol{X}_{i,k}^T\boldsymbol{\beta})+\gamma\sum_{(j,k)\neq(i,l)} s_{i,j}(Y_{j,k}-\boldsymbol{X}_{j,k}^T\boldsymbol{\beta})+\varepsilon_{i,l},$$

where $\boldsymbol{Y}_{-(i,l)}$ denotes all other phenotypic values except $Y_{i,l}$; $\boldsymbol{X}_{i,l}$ and $\boldsymbol{\beta}$ are, respectively, covariates and the corresponding regression coefficients, and thus $\boldsymbol{X}_{i,l}^T\boldsymbol{\beta}$ is the contribution to outcome mean from non-genetic covariates; $\varepsilon_{i,l} \sim$ i.i.d. $N(0,\sigma^2)$; $w(t_{i,k},t_{i,l};\boldsymbol{\eta})$ is the within-subject similarity between $Y_{i,k}$ and $Y_{i,l}$ with parameters $\boldsymbol{\eta}$ playing the role of introducing within-subject correlation between repeated measurements, similar to parameters in a correlation matrix in a GEE framework; $s_{i,j}$ is the genetic similarity between subjects i and j. Possible forms can be $s_{i,j} = \sum_{h=1}^{q}(G_{i,h}-2p_h)(G_{j,h}-2p_h)$ referred to as genetic relationship (GR) (Yang et al. (2011)) where p_h is the population allele frequency of h-th SNP in the region, and the identity-by-state (IBS) similarity: $s_{i,j} = \sum_{h=1}^{q}(2-|G_{i,h}-G_{j,h}|)$. Parameter γ measures the magnitude of the joint association between genetic variants and the phenotype. If none of the genetic variants are associated with the phenotype, the phenotype of subject i will be irrelevant to the phenotypes of others regardless of their proximity in the genetic space, i.e., $\gamma = 0$. On the contrary, a large positive γ indicates a strong spatial dependence or equivalently genetic association. Thus, γ can be interpreted as the magnitude of the joint association between the q genetic variants and the phenotype. Briefly, the conditional autoregressive model relates each observation to others measured on the same subject by within-subject similarity $w(t_{i,k}, t_{i,l}; \boldsymbol{\eta})$, and all other observations (including other measurements on the same subject) in the study by genetic similarity $s_{i,j}$.

According to the factorization theorem of Besag (1974), the conditional model (3.1) uniquely determines a joint distribution of $\boldsymbol{Y}$:

$$\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v}, \boldsymbol{v} \sim N(0, \sigma^2\{\boldsymbol{I} - \boldsymbol{W}(\boldsymbol{\eta}) - \gamma\boldsymbol{S}\}^{-1}), \tag{3.2}$$

where $\boldsymbol{I}$ is an $n \times n$ identity matrix; $\boldsymbol{W}(\boldsymbol{\eta})$ and $\boldsymbol{S}$ are matrices ($n \times n$) composed of $w(t_{i,k}, t_{i,l}; \boldsymbol{\eta})$ and $s_{i,j}$, respectively. Specifically, the within-subject similarity matrix $\boldsymbol{W}(\boldsymbol{\eta})$ is block diagonal with block i ($n_i \times n_i$) corresponding to subject i and the (k, l)-th element of block i is $w(t_{i,k}, t_{i,l}; \boldsymbol{\eta})$ except for diagonal elements of $\boldsymbol{W}(\boldsymbol{\eta})$. The genetic similarity matrix $\boldsymbol{S}$ is composed of $m \times m$ block matrices with dimension $n_i \times n_j, i, j = 1, \ldots, m$, and all elements in the (i, j)-th block are $s_{i,j}$ except for the diagonal elements of $\boldsymbol{S}$. The diagonal elements of $\boldsymbol{W}(\boldsymbol{\eta})$ and $\boldsymbol{S}$ are 0 as in model (3.1) observations are not compared with themselves. To evaluate the joint association of multiple genetic variants with the phenotype we can test the null hypothesis $H_0 : \gamma = 0$ involving a single parameter in the precision matrix (or equivalently in the variance matrix).

With respect to the within-subject similarity, the random field model focuses on how the observations are related, regardless of the direction (past or future) as opposed to transition models which condition each observation only on the past observations. However, they can result in very similar marginal correlation structures such as the first-order auto-regressive (AR1) correlation. Examples of plausible $\boldsymbol{W}(\boldsymbol{\eta})$ are given below.

Example 1. One might assume observations from the same subject to be equally similar and sets $w(t_{i,k}, t_{i,l}; \boldsymbol{\eta}) = \eta$ for $\forall i, k, l$, and in matrix notation, $\boldsymbol{W}(\boldsymbol{\eta}) = \eta \boldsymbol{T}$, where $\boldsymbol{T}$ is a block diagonal matrix with block i, $i = 1, \ldots, m$, an $n_i \times n_i$ matrix with 0's in the diagonal and 1's off-diagonal. Under $H_0 : \gamma = 0$, the corresponding covariance matrix is $\sigma^2(\boldsymbol{I} - \eta \boldsymbol{T})^{-1}$. This specification is equivalent to the usual compound symmetric correlation.

Example 2. One might assume each observation conditionally depends on only the nearest observations before and after it (Markov property): $w(t_{i,k}, t_{i,l}; \boldsymbol{\eta}) = \eta$ if $|k - l| = 1$, and 0 otherwise. This is an approximation of the usual AR1 correlation by ignoring the edge effect (Qu, et al., 2000). Again $\boldsymbol{W}(\boldsymbol{\eta}) = \eta \boldsymbol{T}$ for a block diagonal matrix $\boldsymbol{T}$, where the (k, l)-th element of the i-th block is 1 if $|k - l| = 1$ and 0 otherwise.

In addition, multiple within-subject similarities can be combined for a better working precision matrix, adaptively approximating the underlying structure. Taking $\boldsymbol{W}(\boldsymbol{\eta})$ to be linear in $\boldsymbol{\eta}$, e.g., the two examples given above and their linear combinations, can lead to a rich class to accommodate many commonly used working correlation structures. A similar idea has been studied by Qu et al. (2000) to improve efficiency of estimation over GEE method.

As in the GEE framework, the within-subject similarity matrix $\boldsymbol{W}(\boldsymbol{\eta})$, or equivalently the correlation matrix $\{\boldsymbol{I} - \boldsymbol{W}(\boldsymbol{\eta})\}^{-1}$ under the null, is only a working assumption that is not required to be correct for valid inference. Thus we present our test using a working within-subject similarity matrix that is of the form $\eta \boldsymbol{T}$, as in the two examples, and note the method applies to more general $\boldsymbol{W}(\boldsymbol{\eta})$. For simplicity, the matrix representation of the LGRF model is given by:

$$\boldsymbol{Y}|\boldsymbol{Y}_{-} = \boldsymbol{X}\boldsymbol{\beta} + (\eta \boldsymbol{T} + \gamma \boldsymbol{S})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \tag{3.3}$$

where $\boldsymbol{Y}$ is the n dimensional vector of all observations; $\boldsymbol{Y}|\boldsymbol{Y}_-$ stands for that each observation $Y_{(i,l)}$ is conditional on all other observations $\boldsymbol{Y}_{-(i,l)}$; Matrices $\boldsymbol{T}$ and $\boldsymbol{S}$ have diagonal elements equal to zero, to reflect that the mean of each element of $\boldsymbol{Y}$ only depends on other elements but not on itself; $\boldsymbol{\varepsilon} = (\varepsilon_{1,1}, \ldots, \varepsilon_{1,n_1}, \varepsilon_{2,1}, \ldots)^T$ is the residual vector. Since the genetic similarities are compared across all observations, the model does not have the Markov property, i.e., each observation has finite neighbors, typically assumed in a conditional auto-regressive model in spatial statistics. Thus the regular likelihood-ratio test or score test used in spatial statistics for testing spatial auto-correlation cannot be applied directly. Also, because of the within-subject similarly, the pseudo-likelihood approach developed by He et al. (2014) does not apply. Instead, we propose a set of generalized score type tests.

3.2.2 Association Test under the Longitudinal Genetic Random Field Model

In this subsection we focus on developing a generalized score type test for testing $H_0 : \gamma = 0$ under model (3.3). The inference procedure is developed by treating the within-subject correlation as a working model, leading to a test that is robust to misspecification of the correlation structure. Model (3.3) states, given all other observations, the conditional mean of each observation is linearly related to others, i.e., $E(\boldsymbol{Y}|\boldsymbol{Y}_-) = \boldsymbol{X\beta} + (\eta\boldsymbol{T} + \gamma\boldsymbol{S})(\boldsymbol{Y} - \boldsymbol{X\beta})$. Adopting the similar argument for the usual GEE method (Liang and Zeger (1986)) to our conditional auto-regressive model, we construct the following generalized estimating function:

(3.4)
$$U_\gamma(\boldsymbol{\beta}, \eta, \gamma) = \frac{\partial E(\boldsymbol{Y}|\boldsymbol{Y}_-)^T}{\partial \gamma} \{\boldsymbol{Y} - E(\boldsymbol{Y}|\boldsymbol{Y}_-)\} = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{S}(\boldsymbol{I} - \eta\boldsymbol{T} - \gamma\boldsymbol{S})(\boldsymbol{Y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu} = \boldsymbol{X\beta}$. The estimating equation is quadratic in $\boldsymbol{Y}$ because γ is a coefficient in an auto-regressive model and corresponds to a parameter in the marginal variance as in

(3.2). In the Supplementary Materials (Appendix B) section 1.1, we show that the above estimating function is unbiased in the sense that its expectation is zero under the truth. Therefore, following Boos (1992), we refer to it as a "generalized" score and the score evaluated at $\gamma = 0$, i.e., $U_\gamma(\boldsymbol{\beta}, \eta, 0) = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{S}(\boldsymbol{I} - \eta \boldsymbol{T})(\boldsymbol{Y} - \boldsymbol{\mu})$, can be used to construct a generalized score type test. Due to the unbiasedness, we show that $U_\gamma(\boldsymbol{\beta}, \eta, 0)$ has mean 0 under H_0 and positive mean $\gamma E\{(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{S}^2(\boldsymbol{Y} - \boldsymbol{\mu})\}$ under $H_1 : \gamma > 0$. This rationale leads to constructing a generalized score statistic

$$Q_G = \frac{U_\gamma(\widehat{\boldsymbol{\beta}}, \widehat{\eta}, 0)}{m} = \frac{(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})^T \boldsymbol{S}(\boldsymbol{I} - \widehat{\eta} \boldsymbol{T})(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})}{m} \tag{3.5}$$

and rejecting H_0 when it is sufficiently large. In (3.5), $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\widehat{\eta}$ are estimates under the null hypothesis that $\gamma = 0$. Specifically, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\eta}$ are the solution to the following estimating equations:

$$\begin{cases} U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \eta, 0) = \boldsymbol{X}^T(\boldsymbol{I} - \eta \boldsymbol{T})(\boldsymbol{Y} - \boldsymbol{\mu}) = 0 \\ U_\eta(\boldsymbol{\beta}, \eta, 0) = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{T}(\boldsymbol{I} - \eta \boldsymbol{T})(\boldsymbol{Y} - \boldsymbol{\mu}) = 0. \end{cases}$$

The first equation is the usual estimating equation for estimating $\boldsymbol{\beta}$ in GEE based on the the joint distribution (2) as $\boldsymbol{I} - \eta \boldsymbol{T}$ is proportional to the inverse of a working correlation matrix under H_0. The second equation is derived by considering the estimating function $\frac{\partial E(\boldsymbol{Y}|\boldsymbol{Y}_-)}{\partial \eta}^T \{\boldsymbol{Y} - E(\boldsymbol{Y}|\boldsymbol{Y}_-)\}$ under H_0. It is worth noting that the second estimating equation is linear in η. Thus the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\eta}$ can be calculated by iteratively solving linear equations. This property remains when we linearly combine multiple within-subject similarities, leading to an efficient way to estimate the correlation structure.

We derive an asymptotically equivalent representation of Q_G under H_0 and show that this representation allows us to achieve theoretical protection against the misspecification of within-subject correlation as well as facilitating computationally efficient implementation suitable for large-scale studies. Specifically, we show in Supplementary Materials

(Appendix B) sections 1.2 and 1.3 that for all the genetic similarity metrics introduced previously, under H_0, Q_G can be represented as

$$Q_G = \frac{1}{2m}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}})^T \begin{pmatrix} \boldsymbol{O}_{dq} & \boldsymbol{I}_{dq} \\ \boldsymbol{I}_{dq} & \boldsymbol{O}_{dq} \end{pmatrix} \boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}}) + c + o_p(1),$$

where η_0 is the true parameter under H_0; $\boldsymbol{I}_{dq}$ is a $dq \times dq$ identity matrix and $\boldsymbol{O}_{dq}$ is a zero matrix; $\boldsymbol{R}_1(\eta, \boldsymbol{\beta}) = \widetilde{\boldsymbol{Z}}(\eta)^T(\boldsymbol{Y} - \boldsymbol{\mu})$ and $\widetilde{\boldsymbol{Z}}(\eta) = \{(\boldsymbol{I} - \eta\boldsymbol{T})\boldsymbol{Z}, \boldsymbol{Z}\}$; $\boldsymbol{Z}$ is an $n \times dq$ matrix for some integer d, and c is a constant. The exact form or value of $\boldsymbol{Z}$, d and c depend on the chosen genetic similarity and the details are given in the Supplementary Materials (Appendix B) sections 1.2 and 1.4. For example, for GR similarity, $\boldsymbol{Z}, (n \times q),$ is the centered genotype matrix, i.e., each column of the genotype matrix $\boldsymbol{G}$ is now centered by the genotype population mean $2p_h$. Note that $\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{m}\{\widetilde{\boldsymbol{Z}}_i(\eta_0)^T(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i)\}$, which is a summation of m terms each with expectation zero under the null regardless of the specified working correlation structure. Therefore, the summand is an unbiased estimating function for $\boldsymbol{\beta}$, and according to the theory of M-estimators (Stefanski and Boos (2002)), $\frac{1}{\sqrt{m}}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}})$ is asymptotically normal with a covariance matrix that can be robustly estimated by some sandwich covariance estimates, leading to robustness to misspecification of working correlation.

In Results 2 and 3 of Supplementary Materials (Appendix B), using the theory of M-estimation as well as distributions for quadratic forms, we show that Q_G has an asymptotic distribution

$$\frac{1}{2}\sum_{k=1}^{2dq}\lambda_k\chi_k^2 + c$$

under H_0, where c is a constant which does not affect the inference; χ_k^2's are i.i.d. Chi-square distributions with degree of freedom one; λ_k are eigenvalues of a $2dq \times 2dq$ matrix

$$\begin{pmatrix} \boldsymbol{O}_{dq} & \boldsymbol{I}_{dq} \\ \boldsymbol{I}_{dq} & \boldsymbol{O}_{dq} \end{pmatrix} \boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ can be consistently estimated by a sandwich covariance estimate $\widehat{\boldsymbol{\Sigma}}$, defined in Result 2 of the Supplementary Materials (Appendix B). Moreover, the null distribution of Q_G only depends on the eigen-values of a $2dq \times 2dq$ matrix. As the number of variants in a target gene q is relatively small, it is computationally efficient and hence suitable for large scale GWAS. To obtain the p-value, Davies' method (Davies (1980)) can be used as a computationally efficient way to analytically calculate the tail probability of a mixture of chi-squares by inverting the corresponding characteristic function.

3.2.3 Testing for the Joint Effect of Gene and Gene-time Interaction

As in a regression framework interaction effect is typically modeled using new variables defined as the product of two interacting factors, similarly, we can define interaction terms, $\boldsymbol{G}_i t_{i,l} = (G_{i,1}t_{i,l}, G_{i,2}t_{i,l}, \ldots, G_{i,q}t_{i,l})^T$, and treat them the same way as $\boldsymbol{G}_i$. Therefore the modified LGRF is given by:

$$\begin{aligned} Y_{i,l}|\boldsymbol{Y}_{-(i,l)} &= \boldsymbol{X}_{i,l}^T\boldsymbol{\beta} + \sum_{k\neq l} w(t_{i,k}, t_{i,l}; \boldsymbol{\eta})(Y_{i,k} - \boldsymbol{X}_{i,k}^T\boldsymbol{\beta}) + \gamma_1 \sum_{(j,k)\neq(i,l)} s_{i,j}(Y_{j,k} - \boldsymbol{X}_{j,k}^T\boldsymbol{\beta}) \\ &+ \gamma_2 \sum_{(j,k)\neq(i,l)} s_{il,jk}^{GT}(Y_{j,k} - \boldsymbol{X}_{j,k}^T\boldsymbol{\beta}) + \varepsilon_{i,l}, \end{aligned}$$

where $s_{il,jk}^{GT}$ is the similarity generated by gene-time interaction terms, similar to the genetic similarity; and γ_1 and γ_2 represent the genotype main effect and gene-time interaction effect, respectively. The IBS similarity is not suitable for the interaction terms because it is specifically designed for genetic variants/imputed dosage lying between 0 and 2. In the spirit of genetic relationship similarity, we define $s_{il,jk}^{GT} = \psi(\boldsymbol{G}_i t_{i,l}, \boldsymbol{G}_j t_{j,k}) = \sum_{h=1}^{q}(G_{i,h}t_{i,l} - \overline{G_h t})(G_{j,h}t_{j,k} - \overline{G_h t})$, where $\overline{G_h t} = \frac{1}{n}\sum_{(i,l)} G_{i,h}t_{i,l}$. Considering a working within-subject similarity matrix $\eta\boldsymbol{T}$ as before, in matrix form the model is written as

$$\boldsymbol{Y}|\boldsymbol{Y}_{-} = \boldsymbol{X}\boldsymbol{\beta} + (\eta\boldsymbol{T} + \gamma_1\boldsymbol{S} + \gamma_2\boldsymbol{S}_{GT})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \tag{3.6}$$

where $\boldsymbol{S}_{GT}$ is the similarity matrix of the interaction terms with the (l,k)-th element of the (i,j)-th block $(n_i \times n_j)$ equal to $s^{GT}_{il,jk}$ except for the diagonal of $\boldsymbol{S}_{GT}$. Under this model, we can evaluate the joint effect of gene and gene-time interaction by testing $H_0^J : \gamma_1 = \gamma_2 = 0$.

Denoting $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T$, following previous development, we construct two estimating function with respect to γ_1 and γ_2:

$$\begin{cases} U_{\gamma_1}(\boldsymbol{\beta}, \eta, \boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{S}(\boldsymbol{I} - \eta\boldsymbol{T} - \gamma_1\boldsymbol{S} - \gamma_2\boldsymbol{S}_{GT})(\boldsymbol{Y} - \boldsymbol{\mu}) \\ U_{\gamma_2}(\boldsymbol{\beta}, \eta, \boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{S}_{GT}(\boldsymbol{I} - \eta\boldsymbol{T} - \gamma_1\boldsymbol{S} - \gamma_2\boldsymbol{S}_{GT})(\boldsymbol{Y} - \boldsymbol{\mu}). \end{cases}$$

As before, evaluating the corresponding estimating functions at $H_0^J : \gamma_1 = \gamma_2 = 0$ leads to the following generalized score statistics

$$\begin{cases} Q_G = U_{\gamma_1}(\hat{\boldsymbol{\beta}}, \hat{\eta}, \boldsymbol{O})/m = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})^T \boldsymbol{S}(\boldsymbol{I} - \eta\boldsymbol{T})(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})/m \\ Q_{GT} = U_{\gamma_2}(\hat{\boldsymbol{\beta}}, \hat{\eta}, \boldsymbol{O})/m = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})^T \boldsymbol{S}_{GT}(\boldsymbol{I} - \eta\boldsymbol{T})(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})/m. \end{cases}$$

We propose to combine these two by:

$$Q_J = \alpha_G Q_G + \alpha_{GT} Q_{GT} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})^T (\alpha_G \boldsymbol{S} + \alpha_{GT}\boldsymbol{S}_{GT})(\boldsymbol{I} - \eta\boldsymbol{T})(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})/m,$$

where $\alpha_G = \sqrt{\frac{v_{GT}^2}{v_{GT}^2 + v_G^2}}$ and $\alpha_{GT} = \sqrt{\frac{v_G^2}{v_{GT}^2 + v_G^2}}$; $v_G^2 = 2\text{tr}(\boldsymbol{S}^2)$ and $v_G^2 = 2\text{tr}(\boldsymbol{S}_{GT}^2)$ are proportional to the variance of U_{γ_1} and U_{γ_2} respectively. Though the choice of weights can be arbitrary depending on the need of assessing marginal or interaction effect, our weights are defined such that $\alpha_G Q_G$ and $\alpha_{GT} Q_{GT}$ have approximately equal variance. Defining $\boldsymbol{Z}_{GT}$ as the centered gene-interaction matrix, i.e., each gene-interaction term $G_{i,h}t_{i,l}$ is centered by the its mean $\overline{G_h t}$, $\widetilde{\boldsymbol{Z}}_J(\eta) = \{\alpha_G^{\frac{1}{2}}(\boldsymbol{I} - \eta\boldsymbol{T})\boldsymbol{Z}, \alpha_{GT}^{\frac{1}{2}}(\boldsymbol{I} - \eta\boldsymbol{T})\boldsymbol{Z}_{GT}, \alpha_G^{\frac{1}{2}}\boldsymbol{Z}, \alpha_{GT}^{\frac{1}{2}}\boldsymbol{Z}_{GT}\}$ and

$\boldsymbol{R}_{J1}(\eta, \boldsymbol{\beta}) = \widetilde{\boldsymbol{Z}}_J(\eta)^T(\boldsymbol{Y} - \boldsymbol{\mu})$, we can rewrite the joint test statistic as a quadratic form:

$$Q_J = \frac{1}{2m}\boldsymbol{R}_{J1}(\eta_0, \widehat{\boldsymbol{\beta}})^T \begin{pmatrix} \boldsymbol{O}_{(d+1)q} & \boldsymbol{I}_{(d+1)q} \\ \boldsymbol{I}_{(d+1)q} & \boldsymbol{O}_{(d+1)q} \end{pmatrix} \boldsymbol{R}_{J1}(\eta_0, \widehat{\boldsymbol{\beta}}) + c_J + o_p(1),$$

where d is a constant depending on the chosen genetic similarity for the marginal genetic effect as in Section 2.4 and c_J is a constant similar to c. Although more complex, Q_J has an identical form as Q_G in Section 2.4. The inference follows directly from previous development and therefore we omit the details. The proposed method does not test the gene-time interaction separately; instead, it improves the power of LGRF test by exploiting the potential interaction effect if exists.

3.3 Application: Multi-Ethnic Study of Atherosclerosis

We refer to the LGRF test for the marginal effect of a gene as LGRF-G and the joint test as LGRF-J. We illustrate the proposed methods using data from the Multi-Ethnic Study of Atherosclerosis (MESA). MESA is a collaborative longitudinal study initiated in July 2000 to investigate the prevalence, correlates, and progression of subclinical cardiovascular disease (CVD) (Bild et al. (2002)). From 2000 to 2007, four examinations of blood pressure (BP) were conducted over 18- to 24-month periods. We aimed to replicate the findings (29 significant SNPs associated with blood pressure) of the International Consortium for Blood Pressure (ICBP) (ICBP (2011)) by a region based analysis. A total of 6361 subjects consisting of 2526 Caucasians (CAU), 1611 African Americans (AFA), 1449 Hispanics (HIS) and 775 Asian of Chinese descent (CHN) with genome-wide genotype data, systolic blood pressure (sBP) and diastolic blood pressure (dBP) outcomes were considered in the current analysis. The longitudinal summaries and characteristics of the study population, descriptive statistics are provided in Supplementary Tables B.8 - B.11. For this analysis, we used SNPs that have been directly genotyped using the Affymetrix

Genome-Wide Human SNP Array 6.0 or imputed as per MESA protocol. Imputation was performed using the IMPUTE 2.1.0 program (Marchini et al. (2007)) in conjunction with HapMap Phase I and II reference panels (CEU+YRI+CHB+JPT, release 22 - NCBI Build 36 for African-, Chinese- and Hispanic-American participants; CEU, release 24 - NCBI Build 36 for European Americans). We selected genomic regions around the 29 index SNPs that have demonstrated significant (p-value $< 10^{-9}$) by the ICBP. Each genomic region was defined according to the following criteria: when the index SNP fell within a gene, we selected all SNPs within the gene +/- 5kb and adopted the gene's name. When the index SNP fell outside of a gene, we selected the index SNP plus all SNPs +/- 50kb and name the region after the index SNP. All SNPs are included in the analysis without any minor allele frequency filters. We applied LGRF-G and LGRF-J using longitudinal outcomes and SKAT using the average value of repeated measures to test the association between each candidate region and BPs (sBP and dBP) for the four ethnic groups separately, adjusting for age, gender, BMI and top two principal components (PCs) to correct for potential within-ethnicity stratification. Since only the first two principal components were associated with either systolic or diastolic blood pressure in any ethnicity at $p < 0.01$ (Supplementary Table B.7), we only included these two principal components as adjustment variables. We adjusted the measured blood pressures for participants taking anti-hypertension medications using the standard procedure of adding 10 mmHg to systolic blood pressure and 5 mmHg to diastolic blood pressure (Cui et al. (2003)). The SKAT was implemented with a linear kernel and equal weights on the SNPs. Based on the p-values of the stratified analysis, a meta-analysis was done by Fisher's method.

We analyzed 29 regions with details summarized in the Supplementary Tables B.12 - B.21. The LGRF-G test results in comparable or smaller p-values than SKAT using average outcomes in most cases. We expect LGRF-J to have higher power than LGRF-G when

Table 3.1: Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data: top two regions associated with systolic blood pressure/diastolic blood pressure. Each cell shows the p-value. CAU: Caucasians; AFA: African Americans; HIS: Hispanics; CHN: Asians of Chinese descent. Meta: Meta-analysis combining the results of four ethnic groups using Fisher's combined probability test. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is compound symmetric. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome.The column "SNPs" shows the total number of typed and imputed SNPs in each ethnic group.

	Systolic Blood Pressure							
	Region Indexed by $rs13082711$				Region Indexed by $rs1378942$			
	SNPs	LGRF-G	LGRF-J	SKAT-Avg.	SNPs	LGRF-G	LGRF-J	SKAT-Avg.
CAU	111	0.0052	0.0078	0.0047	84	0.0019	0.0023	0.0019
AFA	82	0.6750	0.6315	0.6806	70	0.1894	0.2047	0.1929
HIS	82	0.0267	0.0453	0.0307	70	0.5269	0.3446	0.4094
CHN	79	0.0191	0.0496	0.0302	70	0.8798	0.9364	0.8969
Meta	-	0.0009	0.0036	0.0013	-	0.0258	0.0248	0.0222
	Diastolic Blood Pressure							
	Region Indexed by $rs13082711$				$C10orf107$			
	SNPs	LGRF-G	LGRF-J	SKAT-Avg.	SNPs	LGRF-G	LGRF-J	SKAT-Avg.
CAU	111	0.1774	0.1185	0.1704	190	0.0283	0.0412	0.0202
AFA	82	0.0263	0.0222	0.0233	157	0.0129	0.0106	0.0152
HIS	82	0.0086	0.0349	0.0058	157	0.0104	0.0081	0.0234
CHN	79	0.0292	0.0713	0.0308	154	0.5361	0.4998	0.4757
Meta	-	0.0006	0.0024	0.0004	-	0.0010	0.0009	0.0015

there exists gene-time interaction, but lower power when there is no such interaction. In the MESA example, the LGRF-J test has smaller p-values than LGRF-G in relatively few instances (for example association of C10orf107 with diastolic blood pressure in Table 1), but larger p-values than LGRF-G in general. This may indicate that gene-time interaction does not have sufficient contribution to the marginal gene-level association in most cases. Table 3.1 shows the results of the top two associations between sBP/dBP and candidate regions. The top two regions were selected according to the p-values of LGRF-G in meta-analysis using Fisher's combined probability test. The region indexed by $rs13082711$ emerged as the most strongly associated region. The meta-analysis p-values of LGRF-G are 8.69×10^{-4} for sBP and 6.25×10^{-4} for dBP. Another suggestive association identified by LGRF-G that is consistent with the ICBP analysis is between dBP and $C10orf107$ (p-value= 9.71×10^{-4}), and LGRF-J exhibited a smaller p-value for this association (p-value= 8.64×10^{-4}).

3.4 Simulation Studies

We evaluated three classes of methods: (a) the proposed multi-marker tests for longitudinal data: LGRF-G, LGRF-J; (b) a multi-marker test in cross-sectional studies using the average of the repeated measures as a single outcome: SKAT; and (c) single-marker tests for longitudinal outcomes: namely, GEE, adjusted by the Bonferroni correction. Specifically, in LGRF-G, LGRF-J and GEE, a working compound symmetric correlation structure was used, and SKAT was implemented with equal weights on the SNPs. Classes (b) and (c) represent two commonly used strategies in practice as currently no multi-marker tests are available for longitudinal data and the specific method (SKAT and GEE) is chosen to be the representative in each class, recognizing that multiple other alternatives in each class exist. Additional simulation studies with respect to the impact of different genetic similarity measures, further evaluation of the power gain using a longitudinal design, use of LGRF in a meta-analysis, and evaluation of type-I error rate and power at lower significance levels are showed in the Supplementary Tables B.2 - B.7.

For each replicated dataset, subjects were randomly selected from the Caucasian (CAU) ethnic group in MESA, and the variants commonly existing in all four ethnicities (154 SNPs) in genotype region $C10orf107$ are included as the target region. We varied the number of repeated measurements to be 4, 6 and 8, and number of subjects 600, 400 and 300 respectively, keeping total number of observations as 2400. Assuming missing completely at random, we first simulated the complete data, and then a missingness indicator with fixed drop-out rate of 4% at each exam was applied approximating what is observed in the MESA study.

We evaluated the type-I error rate at level $\alpha = 0.05$, 0.01, and 0.001 using 100000 replicates. Data are generated from the model:

$$Y_{i,l} = \alpha_0 t_{i,l} + \epsilon_{i,l}, t_{i,l} = 1, \dots, r, \tag{3.7}$$

where $\alpha_0 = \frac{12}{r}$; r is the number of measurements per subject; $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \dots, \varepsilon_{i,r})^T$ independently follows multivariate normal distribution with four types of covariance matrices:

- Independent (Ind.): $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2_{ind}\boldsymbol{I}_r)$.
- Auto-regressive of order 1 (AR1): $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma}_{AR})$, where $\boldsymbol{\Sigma}_{AR}$ is an $r \times r$ matrix and its (l, k) element is $\rho^{|l-k|}\sigma^2_{AR}$.
- Compound symmetry (CS): $\epsilon_{i,l} = b_i + \epsilon^*_{i,l}$, $\epsilon^*_{i,l} \sim N(0, \sigma^2_{error})$, $b_i \sim N(0, \sigma^2_{CS})$, where $\epsilon^*_{i,l}$ and b_i are independent.
- Mixed model with a random intercept and a random slope (RR): $\epsilon_{i,l} = b_{1i} + b_{2i}t_{i,l}/r + \epsilon^*_{i,l}$, $\epsilon^*_{i,l} \sim N(0, \sigma^2_{error})$, $b_{1i}, b_{2i} \sim N(0, \sigma^2_{RR})$, where $\epsilon^*_{i,l}$, b_{1i} and b_{2i} are independent.

Where σ^2_{ind}=16; $\sigma^2_{AR} = 6$, $\rho = 0.6$; $\sigma^2_{error} = 2.25$; $\sigma^2_{CS} = 2.25$; $\sigma^2_{RR} = 1$. The missingness indicator was then applied to the simulated data with 4% drop-out rate. The empirical type-I error rates are presented in Table B.5. LGRF-G and LGRF-J both have well controlled type-I error rates under all scenarios, even if the true correlation is not the assumed working correlation "CS". The tests also have valid type-I error rates at low α-levels (0.01 and 0.001). The simulation results demonstrate that, consistent with the asymptotic result, the proposed methods are robust to misspecification of within-subject correlation in finite samples. We note that the proposed methods tend to be slightly conservative at lower significance levels (Supplementary Table B.5) due to the use of sandwich estimator as in regular GEE.

In the first set of power simulations, one out of 154 SNPs was randomly set to be causal. We evaluated two distinct scenarios where the effect of the single causal SNP is manifested through: 1. its marginal association with outcome, without any gene-time interaction; 2. its interaction with time (SNP × Time interaction). The data was generated respectively:

$$\text{(3.8)} \qquad \text{1. Gene marginal effect} : Y_{i,l} = \alpha_0 t_{i,l} + \alpha_1 G_i + \epsilon_{i,l}, t_{i,l} = 1, \ldots, r,$$

$$\text{(3.9)} \qquad \text{2. Gene-time interaction} : Y_{i,l} = \alpha_0 t_{i,l} + \alpha_2 G_i t_{i,l} + \epsilon_{i,l}, t_{i,l} = 1, \ldots, r,$$

where G_i is the genotype of subject i for the randomly selected causal SNP; $\alpha_0 = 12/r$, $\alpha_1 = 0.4$ and $\alpha_2 = 0.6/r$; r is the number of measurements per subject. To mimic the real data scenario, α_1 and α_2 were elicited based on fitting single SNP models with and without gene-time interaction to MESA data. We chose a large α_0 in our simulation studies to illustrate the power gain that can be expected from a longitudinal design with strong time trend in the mean outcome levels compared to using the average of repeated measures. We recognize that smaller values of α_0 will lead to smaller power differences.

In the second set of simulations, ten out of 154 were randomly set to be causal each time. Among them, six SNPs have only marginal effects, three have both marginal and interaction effects and the remaining one has only an interaction effect. The true model is of the form:

$$Y_{i,l} = \alpha_0 t_{i,l} + \alpha_1^* \sum_{1 \leq k \leq 9} G_{i,k} + \alpha_2^* \sum_{7 \leq k \leq 10} G_{i,k} t_{i,l} + \epsilon_{i,l}, t_{i,l} = 1, \ldots, r.$$

Where $G_{i,k}$ is the genotype of subject i on the k-th randomly selected causal SNP. The coefficients are proportional to α_1 and α_2: $\alpha_1^* = \alpha_1/10 = 0.04$ and $\alpha_2^* = \alpha_2/10 = 0.06/r$, such that the empirical powers are differentiable.

Two important points are illustrated by this simulation: 1. the advantage of incorporating longitudinal information over using only the average outcome; 2. The use of multi-marker tests over single-marker tests. The proposed multi-marker tests using the longitudinal outcome have larger power than SKAT using the average of outcomes, as the proposed tests use the whole trajectory of longitudinal outcomes as opposed to only information contained in the average. When the number of repeated measurements increases, the power becomes more distinct. Not surprisingly, LGRF-J test has slightly lower power than LGRF-G because gene-time interaction does not exist in these scenarios.

When the causal SNP has only an interaction effect (Table 3.4), the relative performance of the methods using repeated measures compared with the one using average outcome is more distinct. In addition, the joint test LGRF-J is able to further enhance power in these scenarios because it incorporates the gene-time interaction explicitly. We note that the power difference between LGRF and SKAT using average outcome is mainly attributed to the longitudinal design rather than the difference between genetic random field model and SKAT (Supplementary Table B.4).

We also note that the proposed multi-marker tests have larger power than single-marker tests using GEE with Bonferroni correction (Tables 3.3-3.5), consistent with results found in cross-sectional studies where advantages of multi-marker tests over single-marker tests have been demonstrated repeatedly. The advantage in power is more substantial when there are multiple causal SNPs (Table 3.5) than when there is only one causal SNP (Tables 3.3-3.4).

Table 3.2: Type-I Error Rate Corresponding to Different Within-Subject Correlation Structures. Each cell represents the empirical type-I error rate evaluated at α=0.05, 0.01 and 0.001 based on 100000 replicates. The total number of observations is 2,400 and repeated measurements per subject were generated in the same follow-up period according to different correlation structures. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS.

	Type-I Error Rate					
	Four Repeated Measurements (600 Subjects)					
	LGRF-G			LGRF-J		
$\alpha =$	0.05	0.01	0.001	0.05	0.01	0.001
Ind.	0.0495	0.0096	0.0008	0.0493	0.0097	0.0008
CS	0.0493	0.0099	0.0009	0.0491	0.0096	0.0009
AR1	0.0499	0.0097	0.0009	0.0507	0.0097	0.0009
RR	0.0497	0.0094	0.0009	0.0498	0.0096	0.0008
	Six Repeated Measurements (400 Subjects)					
	LGRF-G			LGRF-J		
$\alpha =$	0.05	0.01	0.001	0.05	0.01	0.001
Ind.	0.0501	0.0096	0.0009	0.0501	0.0093	0.0010
CS	0.0501	0.0097	0.0009	0.0488	0.0089	0.0008
AR1	0.0485	0.0093	0.0009	0.0494	0.0097	0.0008
RR	0.0497	0.0096	0.0010	0.0500	0.0095	0.0009
	Eight Repeated Measurements (300 Subjects)					
	LGRF-G			LGRF-J		
$\alpha =$	0.05	0.01	0.001	0.05	0.01	0.001
Ind.	0.0488	0.0091	0.0008	0.0483	0.0091	0.0007
CS	0.0484	0.0092	0.0010	0.0488	0.0090	0.0007
AR1	0.0474	0.0090	0.0008	0.0471	0.0089	0.0009
RR	0.0492	0.0095	0.0008	0.0485	0.0091	0.0008

Table 3.3: Power comparisons when **one** randomly selected SNP is causal and has **a marginal effect**. Each cell represents the empirical power from 500 replicates at level α=0.05. The total number of observations is 2,400 and repeated measurements were recorded in the same follow-up period. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction.

Power: Single SNP Marginal Effect

Four Repeated Measurements (600 Subjects)

	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
Ind.	**0.42**	0.39	0.34	0.26	0.19
CS	**0.53**	0.49	0.43	0.41	0.33
AR1	**0.46**	0.45	0.38	0.32	0.28
RR	**0.58**	0.55	0.46	0.50	0.43

Six Repeated Measurements (400 Subjects)

	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
Ind.	**0.48**	0.47	0.31	0.29	0.26
CS	0.40	**0.41**	0.28	0.28	0.23
AR1	**0.41**	0.38	0.29	0.26	0.21
RR	**0.51**	0.48	0.35	0.42	0.37

Eight Repeated Measurements (300 Subjects)

	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
Ind.	**0.40**	0.39	0.25	0.29	0.23
CS	**0.36**	0.35	0.25	0.22	0.18
AR1	**0.36**	0.36	0.22	0.23	0.21
RR	**0.49**	0.45	0.24	0.34	0.30

Table 3.4: Power comparisons when **one** randomly selected SNP is causal and has **only a gene-time interaction effect**. Each cell represents the empirical power from 500 replicates at level α=0.05. The total number of observations is 2,400 and repeated measurements were recorded in the same follow-up period. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction.

Power: Single SNP×Time Effect					
Four Repeated Measurements (600 Subjects)					
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
Ind.	0.38	**0.39**	0.29	0.21	0.20
CS	0.48	**0.54**	0.36	0.33	0.46
AR1	0.41	**0.49**	0.34	0.27	0.34
RR	0.53	**0.57**	0.39	0.42	0.50
Six Repeated Measurements (400 Subjects)					
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
Ind.	0.38	**0.43**	0.20	0.21	0.23
CS	0.33	**0.44**	0.19	0.17	0.37
AR1	0.31	**0.39**	0.21	0.16	0.21
RR	0.42	**0.50**	0.25	0.27	0.38
Eight Repeated Measurements (300 Subjects)					
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
Ind.	0.32	**0.36**	0.16	0.16	0.19
CS	0.25	**0.36**	0.16	0.12	0.30
AR1	0.25	**0.35**	0.14	0.13	0.16
RR	0.35	**0.44**	0.16	0.16	0.31

Table 3.5: Power comparisons when randomly selected **multiple** SNPs are causal and have **both marginal and interaction effects**. Each cell represents the empirical power from 500 replicates at level α=0.05. The total number of observations is 2,400 and repeated measurements were recorded in the same follow-up period. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction.

Power: Multiple SNPs Combined Effect					
Four Repeated Measurements (600 Subjects)					
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
Ind.	**0.36**	0.36	0.25	0.13	0.09
CS	**0.50**	0.49	0.37	0.19	0.18
AR1	**0.43**	0.42	0.35	0.19	0.17
RR	**0.60**	0.60	0.46	0.36	0.29
Six Repeated Measurements (400 Subjects)					
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
Ind.	**0.37**	0.36	0.21	0.15	0.11
CS	0.33	**0.35**	0.21	0.12	0.10
AR1	**0.32**	0.32	0.22	0.13	0.10
RR	**0.46**	0.43	0.24	0.22	0.15
Eight Repeated Measurements (300 Subjects)					
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
Ind.	0.30	**0.30**	0.17	0.11	0.11
CS	0.27	**0.29**	0.18	0.09	0.11
AR1	0.26	**0.28**	0.14	0.08	0.08
RR	0.40	**0.41**	0.20	0.19	0.15

CHAPTER IV

Set-Based Tests for Gene-Environment Interaction in Longitudinal Studies

4.1 Introduction

Most complex traits have a multifactorial etiology involving the dynamic interplay of genes and environmental exposures over the life course. Studies of gene-environment interaction (GEI) often suffer from single one time measurement of exposure or a crude proxy thereof, without proper characterization of lifetime history of cumulative exposure. Longitudinal studies with time varying measures of outcome and exposure data help with characterizing the temporal features of exposure and outcomes, handling exposure measurement error and often enhance power when compared to a cross-sectional analysis. While environmental factors considered in an epidemiological analysis are often behavioral factors like diet, physical activity, use of tobacco or alcohol, in recent years, there has been an increasing interest in measuring the neighborhood environment that the individual lives in. For example, the MESA neighborhood Study, an ancillary study to the Multi-Ethnic Study of Atherosclerosis (MESA), includes a set of novel time varying measures of healthy food availability and access to recreational facilities. Previous studies have shown that individuals living in neighborhoods with better food and physical activity environments are less likely to develop hypertension (Kaiser et al. (2015)). In the present analysis, we are primarily interested in investigating whether a set of single nucleotide polymorphisms (SNPs) measured in a genome-wide association study modifies the effect of neighborhood exposures on longitudinal measures of blood pressure.

Gene-environment interaction is often statistically assessed by fitting a regression model for the quantitative outcome (Y) by including the main effects and a product between a genetic variant (G) and an environmental exposure (E), adjusting for covariates (X). A typical genome-wide interaction search repeats the test for interaction under this model for millions of SNPs, adjusting for multiple comparison. Although numerous single SNP based analyses for gene-environment interaction have been conducted, relatively few of the findings have been replicated because of various reasons such as: limited statistical power due to the burden of multiple comparison; measurement error and misclassification of exposure; detection of spurious interactions due to not properly adjusting for main effect of E and G (for example due to missing a non-linear terms in a continuous exposure E) (Thomas (2010); Tchetgen Tchetgen and Kraft (2011); Mukherjee et al. (2012); Cornelis et al. (2012); Boonstra et al. (2016)).

To improve power and to reduce the burden of multiple comparison, many genetic association studies have now considered an alternate or supplementary analytic approach towards jointly testing the effect of all SNPs in a biologically defined set, such as a gene, pathway or specific genomic region as opposed to a one-at-a-time single SNP analysis. Aggregation of SNPs is particularly critical for studies of rare variants (Derkach et al. (2014); Basu and Pan (2011)). A number of methods have gained popularity including kernel machine regression methods (Wu et al. (2011)), similarity regression (Tzeng et al. (2011)), sum of squared score test (Pan (2009)) and genetic random field model (He et al. (2014); He et al. (2015)). In the context of testing gene-gene/gene-environment interaction for cross-sectional studies, Tzeng et al. (2011), Li et al. (2012), Lin et al. (2013), Chen et al. (2014), Marceau et al. (2015) and Lin et al. (2016) extended the set-based tests for marginal associations to testing interactions. These papers demonstrated superior power of set-based tests for gene-environment interaction by aggregating signals across multiple SNPs. However, no set-based test for gene-environment interaction has been proposed for longitudinal studies where improved power regarding gene-environment interaction is possible by using longitudinally varying outcome and exposure trajectories.

Most GEI studies consider a linear main effect of E. A growing body of literature has shown that a misspecified main effect of E can lead to type I error inflation in tests for gene-environment interaction, and gene-environment independence in the underlying population plays an important role in reducing the detection of spurious gene-environment interactions (Tchetgen Tchetgen and Kraft (2011); Voorman et al. (2011); Cornelis et al. (2012)). However, the theoretical justification for this result has not been established (VanderWeele et al. (2013)). Also, there is no method proposed for handling misspecified E effect when G and E are dependent, particularly for set-based analysis. For the main effect of G, Lin et al. (2013) pointed out that single SNP analyses for gene-environment interaction can be biased due to ignoring SNPs in the same region that are in linkage disequilibrium (LD) with the tested SNP. Set-based analysis can serve as a potential remedy to this issue, but one practical challenge that is new to deriving set-based tests for GEI is that the null model contains main effects of multiple SNPs and fitting the null model could potentially be problematic when the number of SNPs in a region is large relative to the sample size. The tests can suffer from type I error inflation as the asymptotic distributional properties of the reference test statistic may not hold under such situations.

In this article, we propose a new statistical approach to test for gene-environment interactions with a set of genetic variants and longitudinally measured outcome and exposure data. The test is robust to misspecification of within subject correlation and is substantially more powerful than an analysis that uses subject-specific averages/summaries of outcome and exposure data. We show that the proposed test is robust to the misspecification of E and G main effects under the gene-environment independence condition. We further propose using the method of sieves to flexibly model the main effect of E for improved type I error control when the gene-environment independence condition does not hold, and for better power. We also proposed a weighted principal component analysis (PCA) to remedy the curse of dimensionality when the number of SNPs in the tested set is close to or larger than the sample size. We illustrate the proposed methods by both an analysis of targeted GEI (restricted to genetic regions defined around previous GWAS hits) and an agnostic genome-wide gene-based GEI search, with novel time-varying neighborhood features of the environment as exposure, and blood pressure as the longitudinally measured outcome in MESA. Extensive simulation studies, designed to mimic the data structure of MESA are conducted to assess the operating characteristics of the different methods.

4.2 Application: Multi-Ethnic Study of Atherosclerosis

MESA was initiated in the year 2000 with the goal of investigating the prevalence, correlates and progression of subclinical cardiovascular disease (Bild et al. (2002)). A total of 6360 MESA subjects who consented to genetic analyses, including 2526 European Americans (EUR), 1611 African Americans (AFA), 1448 Hispanics (HIS) and 775 Asian of Chinese descent (CHN), were included in the current analysis. From 2000 to 2007, four examinations were conducted at approximately 1.5-2 year intervals for participants residing at six study sites: New York, New York; Baltimore, Maryland; Forsyth County, North Carolina; Chicago, Illinois; St Paul, Minnesota; and Los Angeles, California. Blood pressure measurements were available at each MESA exam. An ancillary study of MESA, the MESA neighborhood study, collected longitudinal information on neighborhood characteristics in the four examinations, including four time varying measures of healthy food availability and physical activity resources (Moore et al. (2008); Christine et al. (2015)). These neighborhood environments may influence individual diet and exercise levels, and therefore influence risk factors for chronic diseases, e.g. systolic/diastolic blood pressure (Mujahid et al. (2008)).

The four neighborhood measures include two geographic information system (GIS) based measures and two survey based measures: 1. Density of favorable food stores (GIS-based); 2. Density of recreational facilities (GIS-based); 3. Perceived healthy foods availability (survey-based); 4. Perceived walkability (survey-based). The GIS measures were constructed using the National Establishment Time Series (NETS) database from Wall and Associates for 2000 to 2007 on food stores and commercially-available recreational facilities for every ZIP code within a 5 miles radius of MESA participant households. The survey based measures of healthy food availability and walkability were obtained from questionnaires administered to MESA participants and supplementary sample of other community residents. The detailed description of these neighborhood features can be found in section 4.1 the Supplementary Materials (Appendix C). A growing body of literature has suggested that altering these neighborhood environments may foster behavioral changes and may aid in prevention of chronic diseases (Papas et al. (2007); Sallis et al. (2012); Christine et al. (2015)). Our interest lies in understanding whether an individual's genomic profile modifies the effect of neighborhood features on blood pressure.

We conducted both a targeted GEI analysis and a gene-based genome-wide GEI analysis. Our targeted GEI analysis studied 29 candidate genomic regions which were selected around 29 index SNPs that are significantly associated with blood pressures (p-value $< 10^{-9}$) by the International Consortium for Blood Pressure Genome-Wide Association Studies, ICBP (2011). The criteria of determining each genomic region is same as He et al. (2015): when the index SNP fell within a gene, we selected all SNPs within the gene +/- 5kb and adopted the gene's name to label the region. When the index SNP fell outside of a gene, we selected the index SNP plus all SNPs +/- 50kb and name the region after the index SNP. Number of SNPs in these regions ranges from 10 to 840 SNPs. Our genome-wide gene-based analysis studied 24743 protein coding genes +/- 5kb defined by the UCSC genome browser (Karolchik et al. (2003)). The SNPs in the regions were directly genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 or imputed as per MESA protocol. Imputation was performed using the IMPUTE 2.1.0 program by Marchini et al. (2007) in conjunction with HapMap Phase I and II reference panels (CEU+YRI+CHB+JPT, release 22 - NCBI Build 36 for African-, Chinese- and Hispanic-American participants; CEU, release 24 - NCBI Build 36 for European Americans). All common and rare variants are included in our analysis without any minor allele frequency filters.

4.3 Model and Inference

Consider a study population with m independent subjects where the i-th subject has n_i longitudinal observations, $n = \sum_{i=1}^{m} n_i$. When $n_i = 1$ for all $1 \leq i \leq m$, this corresponds to a cross-sectional study. Let $Y_{i,j}$ be the quantitative outcome value, $\boldsymbol{X}_{i,j} = (X_{i,j}^1, ..., X_{i,j}^p)^T$ be the p covariates which can include age, gender, education, etc., $E_{i,j}$ be the environmental exposures for the j-th observations on the i-th subject measured at time $t_{i,j}$; $\bar{\boldsymbol{G}}_i = (G_i^1, ..., G_i^q)^T$ be the q time-invariant genetic variants in the target region, where $G_i^k \in \{0, 1, 2\}$. We define $\boldsymbol{Y}_i = (Y_{i,1}, ..., Y_{i,n_i})^T$ as a vector of all observations and $\boldsymbol{G}_i = (\bar{\boldsymbol{G}}_i, ..., \bar{\boldsymbol{G}}_i)^T$ as an $n_i \times q$ matrix of genetic variants where $\bar{\boldsymbol{G}}_i$ is repeated n_i times; $\boldsymbol{X}_i$, $\boldsymbol{E}_i$ are defined as the matrix forms of covariates and environmental exposure similarly. We are interested in the statistical interaction between $E_{i,j}$ and $\bar{\boldsymbol{G}}_i$ on outcome $Y_{i,j}$, adjusting for $\boldsymbol{X}_{i,j}$ in addition to the main effect of $E_{i,j}$ and $\bar{\boldsymbol{G}}_i$. $E_{i,j}$ can be a summary statistic of measures of environmental exposure prior or up to exam j if the investigators believe the outcome not only depends on the current values of exposure but also the previous exposure history. For example, $E_{i,j}$ can be the cumulative average of repeated exposure measures up to exam j. The statistical interaction between the environmental exposure and the k-th genetic variant is characterized by $E_{i,j}G_i^k$. We define $E_{i,j} * \bar{\boldsymbol{G}}_i = (E_{i,j}G_i^1, ..., E_{i,j}G_i^q)^T$ and its matrix form is denoted by $\boldsymbol{E}_i * \boldsymbol{G}_i$, an $n \times q$ matrix.

One popular approach for analyzing longitudinal genetic data is a single SNP analysis, repeated for each of the G_i^k separately, $k = 1, ..., q$, based on a generalized estimating equation (GEE) approach,

$$E(Y_{i,j}|\boldsymbol{X}_i, \boldsymbol{E}_i, G_i^k) = \boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_X + E_{i,j}\beta_E + G_i^k\beta_{G,k} + E_{i,j}G_i^k\gamma_k,$$

where $\boldsymbol{\beta}_X = (\beta_{X,1}, ..., \beta_{X,p})^T$, β_E and $\beta_{G,k}$ are the coefficients for covariates, main effect of exposure and the k-th SNP respectively; γ_k is the gene-environment interaction parameter of interest. Both the main effects $(\boldsymbol{\beta}_X, \beta_E, \beta_{G,k})$ and the interaction effect γ_k are modeled as fixed effects. The null hypothesis is $H_0 : \gamma_k = 0$. To extend it to a set-based analysis, a natural multivariate model includes all SNPs in the same region simultaneously,

$$\mu_{i,j} = E(Y_{i,j}|\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) = \boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_X + E_{i,j}\beta_E + \bar{\boldsymbol{G}}_i^T\boldsymbol{\beta}_G + (E_{i,j} * \bar{\boldsymbol{G}}_i)^T\boldsymbol{\gamma}, \tag{4.1}$$

where $\boldsymbol{\beta}_G = (\beta_{G,1}, ..., \beta_{G,q})^T$; $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_q)^T$. The null hypothesis jointly tests the entire interaction vector of length q, namely, $H_0 : \boldsymbol{\gamma} = 0$. The working covariance matrix of $\boldsymbol{Y}_i$ is denoted as $\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})$, which is of size $n_i \times n_i$ and depends on a vector of parameters $\boldsymbol{\zeta}$. For cross-sectional studies, Lin et al. (2013) considered $\boldsymbol{\beta}_G$ as fixed effects and assumed that each coefficient γ_k follows i.i.d $N(0, \tau^2)$ and proposed a variance component score test for $H_0 : \tau^2 = 0$. Instead of the mixed effect model, we propose a GEE approach based on the unified fixed effect model (4.1), where the parameters have a more natural interpretation.

The classical approach for testing $H_0 : \boldsymbol{\gamma} = 0$ is a q-degree of freedom likelihood ratio/wald/score test. However, Goeman et al. (2006) showed the power of such tests tend to diminish rapidly when the dimensionality q is large, which is common when the region considered consists of hundreds of variants. To address this, we develop a generalized score type test that can exploit the LD among the SNPs to reduce the test degrees of freedom under model (4.1). The score vector from model (4.1) with respect to $\boldsymbol{\gamma}$ is:

$$S_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}) = \sum_{i=1}^{m} S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma}) = \sum_{i=1}^{m} (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

where $\boldsymbol{\mu}_i = (\mu_{i,1}, ..., \mu_{i,n_i})^T$. By M-estimation theory, the score statistic $\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$ asymptotically follows a multivariate normal distribution with mean zero and covariance $\boldsymbol{\Sigma}$ under H_0, where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\zeta}}$ are the estimators under $H_0 : \boldsymbol{\gamma} = 0$ obtained by using the usual GEE proposed by Liang and Zeger (1986). Each element $\frac{1}{\sqrt{m}} S^k_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$ follows an asymptotic normal distribution with mean zero. The classical score test summarizes the vector $\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$ into a scalar by considering $\frac{1}{m} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)^T \widehat{\boldsymbol{\Sigma}}^{-1} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$ where $\widehat{\boldsymbol{\Sigma}}$ is an estimator of $\boldsymbol{\Sigma}$. In this case, the test statistic follows a chi-square distribution with q degrees of freedom, i.e., a sum of q squared independent normal random variables. This approach involves the inversion of $\widehat{\boldsymbol{\Sigma}}$, which is not stable when q is large relative to m, and cannot be applied to scenarios when $q > m$. To address this, we define a test statistic Q for testing $H_0 : \boldsymbol{\gamma} = 0$ by aggregating the score statistics in a different way,

$$Q = \frac{1}{m} S^T_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) = \frac{1}{m} \sum_{k=1}^{q} \{S^k_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)\}^2,$$

where $S^k_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$ corresponds to the k-th interaction term. The statistic can be understood as the overall deviation from 0 of all score statistics where each of them measures the strength of a specific interaction effect. Let $S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma})$ denote the score vector with respect to $\boldsymbol{\beta}$.

Result 3.1: Under model (4.1) and $H_0 : \boldsymbol{\gamma} = 0$, if q is fixed and $m \to \infty$, Q is asymptotically distributed as

$$\sum_{k=1}^{q} \lambda_k \chi_k^2 \tag{4.2}$$

where χ_k^2s are i.i.d. Chi-square distributions with degree of freedom one; $\lambda_1 \geq \ldots \geq \lambda_q$ are the eigen-values of $\boldsymbol{\Sigma}$ and can be estimated by $\{\widehat{\lambda}_k\}_{1 \leq k \leq q}$,

$$\max_{1 \leq k \leq q} |\widehat{\lambda}_k - \lambda_k| = o_p(1), \quad m \to \infty;$$

$\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_q$ are the ordered eigen-values of $\widehat{\boldsymbol{\Sigma}}$. Specifically, $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{A}}^T$,

$$\widehat{\boldsymbol{A}} = \{\boldsymbol{I}_q, -[\sum_{i=1}^{m}(\boldsymbol{E}_i{*}\boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)][\sum_{i=1}^{m}(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)]^{-1}\},$$

$$\widehat{\boldsymbol{D}} = \frac{1}{m-p-q-1}\sum_{i=1}^{m} S_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) S_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)^T, S_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) = [S_{\boldsymbol{\gamma},i}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)^T, S_{\boldsymbol{\beta},i}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)^T]^T.$$

Result 3.1 shows the asymptotic behavior of the test statistic Q as m goes to infinity. The proof is given in the supplemental materials. The variance component test proposed by Lin et al. (2013) also follows a similar weighted summation of chi-square distributions, but their weights are estimated using a model based inference. Instead, we estimate the weights using the "sandwich estimators". The empirical estimated weights make the test robust against misspecification of within-subject correlation, which is a desirable property in longitudinal studies with repeated measurements. This sandwich estimation also plays a role in reducing spurious gene-environment interactions caused by potential main effect misspecification of E when G and E are independent, as observed by Voorman et al. (2011) and Cornelis et al. (2012). The rigorous result that explains these observations will be left to the next section.

The proposed test statistic belongs to the class of quadratic test statistics of the form $Q = \boldsymbol{S}^T \boldsymbol{A} \boldsymbol{S}$ as described in Derkach et al. (2014), where $\boldsymbol{S}$ is the score vector. Other examples of test statistics which belong to this class include the ones used in the methods rareGE (Chen et al. (2014)), iSKAT (Lin et al. (2013); Lin et al. (2016)) and the classical q d.f. score test. For our proposed test, rareGE and iSKAT, $\boldsymbol{A}$ equals $\boldsymbol{I}$. For the classical score test, $\boldsymbol{A}$ equals $\widehat{\boldsymbol{\Sigma}}^{-1}$ where $\widehat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix of $\boldsymbol{S}$. Since SNPs in a region can be strongly correlated due to linkage disequilibrium, many eigen-values of $\boldsymbol{\Sigma}$ are close to 0, and the effective test degrees of freedom is less than q. Therefore the proposed test implicitly reduces the test degrees of freedom compared to the classical score test. It is worth noting that the power of a test not only depends on the test degrees of freedom, but also the non-centrality parameter. Since both the effective test degrees of freedom and non-centrality parameter may change across various scenarios, there is no theoretical result for a uniformly optimal choice for constructing a test statistic achieving the highest power in the class of quadratic test statistics. However, many empirical studies have demonstrated the tests with $\boldsymbol{A} = \boldsymbol{I}$, such as rareGE, iSKAT and our proposed test, has superior power than classical score test in genetic association studies (Wu et al. (2010); Tzeng et al. (2011); He et al. (2014)). Basu and Pan (2011) also pointed out that these tests can be regarded as modified score test by ignoring the non-diagonal elements of $\boldsymbol{A}$, which is known to be advantageous for high-dimensional data.

RareGE, iSKAT and our test statistic differ in three aspects, even for cross-sectional data. First, they adjust for the main effect of G differently. RareGE assumes the coefficients $\boldsymbol{\beta}_G \sim N(0, \tau_1\boldsymbol{I})$, iSKAT uses ridge regression, our method uses the weighted PCA approach. Second, each of these methods defines the score vector $\boldsymbol{S}$ differently. They all consider a score vector of the form $\boldsymbol{S} = \boldsymbol{G}^T\boldsymbol{V}^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})$, but rareGE defines $\boldsymbol{V}(\widehat{\boldsymbol{\zeta}}) = \hat{\tau}_1\boldsymbol{G}\boldsymbol{G}^T + \hat{\sigma}^2\boldsymbol{I}$, $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_X + \boldsymbol{E}\hat{\beta}_E$; both iSKAT and our test defines $\boldsymbol{V}(\widehat{\boldsymbol{\zeta}}) = \hat{\sigma}^2\boldsymbol{I}$, $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_X + \boldsymbol{E}\hat{\beta}_E + \boldsymbol{G}\widehat{\boldsymbol{\beta}}_G$. Third, they estimate the covariance of $\boldsymbol{S}$ differently. Both rareGE and iSKAT use model based inference, our test uses robust variance estimation.

In addition, we note that rareGE and iSKAT were proposed only for cross-sectional data, whereas our test is applicable to both cross-sectional and longitudinal studies. Moreover, our method proposes flexible modeling of main effects of G & E (Section 4.4), leading to improved power and Type 1 error properties under misspecification of the main effect model.

4.4 Main Effect Adjustment

So far, we have discussed inference under a correctly specified main effect model under H_0. Unlike set-based tests for genetic association, set-based tests for gene-environment interaction face the unique challenge of having a potentially misspecified and high-dimensional null model. In this section, we consider potential strategies when the main effect of E may be misspecified and the dimension of G, namely q, is large relative to m. A key step for implementing the proposed generalized score type test is fitting the following main effect model under the null hypothesis

$$\mu_{i,j} = E_{H_0}(Y_{i,j}|\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) = \boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_X + E_{i,j}\beta_E + \bar{\boldsymbol{G}}_i^T\boldsymbol{\beta}_G.$$

There are two challenges with respect to this step. First, a misspecified main effect of $E_{i,j}$ can lead to a biased score ($E_{H_0}[S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)] \neq 0$) and severe type I error inflation. This may happen when the underlying main effect of the environmental exposure is nonlinear but a linear model is specified. Second, the dimension of $\bar{\boldsymbol{G}}_i$ can be large relative to the sample size, such as the $MECOM$ region in MESA which includes 821 SNPs but the Chinese Americans only have 775 subjects. The estimates of $\boldsymbol{\beta} = (\boldsymbol{\beta}_X^T, \beta_E, \boldsymbol{\beta}_G^T)^T$ are not consistent and the approximation to the asymptotic distribution of Q as presented in Result 3.1 does not hold anymore. To address these challenges, we first ensure the robustness of the proposed test to main effect misspecification by exploiting the gene-environment independence condition, then develop methods to handle the main effect misspecification of E and high-dimensionality of G when the gene-environment independence condition does not hold.

4.4.1 Gene-environment independence condition

Gene-environment independence plays a crucial role in the main effect adjustment. We show in Result 4.1 that the test proposed in Section 4.3 will be robust to main effect misspecification under the gene-environment independence condition, by centering $\boldsymbol{E}_i$ and $\boldsymbol{G}_i$ using weighted average as described in section 1.2 of the Supplementary Materials (Appendix C).

Result 4.1: If the following two assumptions hold:

C1. $\boldsymbol{X}_i$ can be separated as $(\boldsymbol{X}_i^E, \boldsymbol{X}_i^G)$ where $(\boldsymbol{X}_i^E, \boldsymbol{E}_i)$ is independent of $\boldsymbol{G}_i$ and $(\boldsymbol{X}_i^G, \boldsymbol{G}_i)$ is independent of $\boldsymbol{E}_i$,

C2. $\mathrm{cov}(\boldsymbol{X}_{i,l}^G, \boldsymbol{G}_i)$ is time invariant,

then the expectation of the score vector equals zero, i.e., $E_{H_0}[S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)] = 0$, regardless of the main effect model of E and G when $\boldsymbol{E}_i$ and $\boldsymbol{G}_i$ are centered appropriately.

Condition C1 can be seen as the more commonly used condition of gene-environment independence with additional requirement on the covariates $\boldsymbol{X}_i$. For instance, time and age are likely to be correlated with the time varying environmental exposure but independent of the time invariant SNPs. It reduces to the gene-environment independence condition in the special case of no covariates. Condition C2 is specifically for longitudinal studies, and it always holds for cross-sectional studies. It is also satisfied in the special case when $\boldsymbol{X}_i^G$ is time invariant, which is common in a genetic study, e.g. when $\boldsymbol{X}_i^G$ consists of the leading principal components to control for population stratification. The weighted average used to center E and G are proposed to take into account of the within-subject correlation among observations on the same subject (see section 1.2 of the Supplementary Materials

(Appendix C)). For cross-sectional studies, this approach reduces to simply centering E and G by the usual average.

Under C1 and C2, the proposed test is robust to a misspecified main effect model, if $\mathbf{\Sigma}$ is estimated using the sandwich covariance estimator and E and G are weighted and centered. This is because the score statistic $\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$ will asymptotically follow a mean zero multivariate normal distribution, whose covariance matrix is empirically estimated by sandwich estimators. Therefore the asymptotic distribution of Q, as a function of $\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$, can be correctly estimated. Under C1 and C2, this result shows using a linear model for E is sufficient for controlling type I error rate regardless of the true functional form of the main effect of E. The problem of inconsistency due to high-dimensionality of G can be simply solved by excluding the main effects of all SNPs in the model. However, these strategies are not adequate, especially when C1 and C2 are violated. We further develop methods for main effect adjustment of E and G in the subsequent sections that are appropriate under violations of C1 and C2.

This result also explains the findings in Voorman et al. (2011) and Cornelis et al. (2012), where the authors showed that using sandwich estimators can reduce the detection of spurious gene-environment interactions in cross-sectional studies. Specifically, the simulation studies conducted by Voorman et al. (2011) did not observe any type I error inflation under misspecification of main effect of E when a sandwich estimator was used, because no association between G and E was simulated; The genome-wide analysis for gene-environment interactions conducted by Cornelis et al. (2012) used QQ-plots to show that using a sandwich estimator can reduce the type I error inflation. This is likely due to the fact that a vast majority of the SNPs are usually not correlated with the environmental exposure. Using sandwich estimators for variance will eliminate the inflation for these SNPs as gene-environment independence is effectively true in these situations.

4.4.2 Main effect misspecification of E

Most GEI studies consider a linear main effects model as described in (4.1). When C1 and C2 do not hold, ignoring a nonlinear main effect can result in a biased score function and lead to severe type I error inflation. Even if C1 and C2 hold and type I error is not a concern, a misspecified main effect model for E can significantly reduce power for testing interaction. Examples include the cases when the main effect of E has a quadratic effect, or E is a log-transformed exposure but the true effect is on the original scale. In this subsection, we make further effort to control the bias in the scores due to a misspecified main effect of E when C1 and C2 do not hold, and improve the power. Since the true main effect $h_E(\cdot)$ is unknown, we propose to approximate it non-parametrically by the method of "sieves": expand $h_E(\cdot)$ by a sequence of finite dimensional models Φ_U (sieves), then allow the model complexity U to grow slowly with the sample size (Grenander (1981)). Numerous sieve estimators have been proposed such as the polynomial sieves and the spline sieves:

$$\Phi_U^P = \{h_{E,U} : h_{E,U}(x, \boldsymbol{\beta}_E) = \sum_{u=1}^{U} x^u \beta_{E,u}\}; \quad \Phi_U^S = \{h_{E,U} : h_{E,U}(x, \boldsymbol{\beta}_E) = \sum_{u=1}^{U} B_u^U(x) \beta_{E,u}\},$$

where $B_u^U(\cdot)$ is the u-th spline basis function. So the function $h_E(\cdot)$ can be approximated by a series of sieves. The uniform convergence rate of $h_{E,U}(x, \widehat{\boldsymbol{\beta}}_E)$ as $m \to \infty$ depends on the smoothness of $h_E(x)$. The details of asymptotic results can be found in Newey (1997).

The main effect model based on the sieve representation can be written as

$$\mu_{i,j} = E_{H_0}(Y_{i,j} | \boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) = \boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_X + h_{E,U}(E_{i,j}; \boldsymbol{\beta}_E) + \bar{\boldsymbol{G}}_i^T \boldsymbol{\beta}_G, \tag{4.3}$$

where $h_{E,U}(\cdot)$ is a finite dimensional model using spline/polynomial sieves. Result 4.2 shows that, under C1 and C2, a test for gene-environment interaction based on a main effect model (4.3) will be asymptotically equivalent to using the true model. Thus the test not only has correct type I error rate, but also is as powerful as using the true model.

Result 4.2: If C1 and C2 hold and $h_{E,U}(x;\widehat{\boldsymbol{\beta}}_E)$ uniformly converges to $h_E(x)$ for $\forall\ x$ as $m \rightarrow \infty$,

$$\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^0) + o_p(1).$$

where $\boldsymbol{\mu}_i^0$ is the stacked vector of conditional means in (4.3) with true main effect $h_E(\cdot)$.

This includes a scenario where U is larger than the underlying model complexity. For example, if the underlying main effect of E is linear but we model it using cubic-spline sieves with $U > 1$, the test will be asymptotically equivalent to a linear model and will not be less powerful under C1 and C2. When C1 and C2 do not hold, introducing unnecessary model complexity can reduce the power. However, we note that the proportion of total variation of an exposure explained by a single genomic region is usually not expected to be very large. With this weak dependency, our simulation studies demonstrate that type I error inflation due to main effect misspecification of E can be severe, but the power loss due to using more complex model is negligible (Table 4.1). In summary, flexibly modeling the main effect of E does not substaintially hurt power for tests of gene-environment interaction, and greatly helps in controlling type I error rate. This is a very important observation for practice. However, we note that this is different from using more flexible models for the GEI terms in the alternative hypothesis, which certainly entail substantial loss of power.

Result 4.2 also helps to choose the model complexity U, which plays a crucial role in the method of sieves. The common criteria include cross-validation that minimizes the integrated mean square error, the Mallows criterion by Mallows (1973), the Akaike information criterion described in Akaike (1998) and the Bayesian information criterion by Schwarz (1978). Although these methods are still reasonable, Result 4.2 indicates that the ideal criteria for main effect adjustment can be different, because the primary focus is to test another set of variables (the interactions terms). Based on Result 4.2, a larger U allowed by sample size is recommended for better controlling type I error rate and will not hurt power. In this paper, we specifically illustrate the proposed test with a sufficiently rich main effect model for E with $U = m^{\frac{1}{2}}$. The choice of U is driven by existing results that ensure the asymptotic estimation of the coefficients by GEE is reliable. The detailed discussion can be found in Wang (2011).

4.4.3 High-dimensionality of G

When a large genomic region is considered in a set-based analysis, the number of parameters can be large relative to the sample size. The top panel in Supplementary Figure C.1 shows an example in MESA where region $MECOM$ includes 821 SNPs but the Chinese Americans only have 775 subjects. When C1 and C2 do not hold, the main effect of G cannot be ignored because its confounding effect can lead to bias and type I error inflation. Lin et al. (2013) proposed to use ridge regression for handling the main effect of G, but their test is still based on the assumption that q is fixed and $m \to \infty$, same as the method presented in Section 4.3. These methods work well when the dimension of G is moderate, but suffer from severe type I error inflation when the number of SNPs is close to or larger than the sample size (Table 4.2; Supplementary Table C.1). This is a curse of dimensionality and some form of dimension reduction in the G space is needed. In this subsection, we make further effort to deal with both the high-dimensionality and the confounding effect of G.

To deal with the high-dimensionality of G, one natural choice is taking advantage of the LD structure in genetic regions, and use some form of PCA. The first panel in Supplementary Figure C.1 shows a typical genome region that contains several LD blocks and SNPs within each block are correlated. Therefore eigen-values corresponding to the principal components (PC) decrease to zero very quickly as a function of the leading number of components (Supplementary Figure C.1). This enables us to use a small number of PCs to explain most variation in G. A standard PCA results in orthogonal components $\{P_i^s\}_{1\leq s\leq q}$ ranked by the corresponding eigen-values $\kappa_1 \geq \ldots \geq \kappa_q$, $E(P_i^s) = 0$, $var(P_i^s) = \kappa_s$. Each component is a linear combination of $\{G_i^k\}_{1\leq k\leq q}$. We usually fit the leading PCs:

$$\mu_{i,j} = E_{H_0}(Y_{i,j}|\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) = \boldsymbol{X}_{i,j}^T\boldsymbol{\beta}_X + E_{i,j}\beta_E + \sum_{s=1}^{S} P_i^s\beta_{P,s}, \tag{4.4}$$

where $1 \leq S \leq q$. The PCA approach with a well chosen S is a plausible remedy for the curse of dimensionality, but not ideal for adjusting the confounding effect of G because there can be low-rank PCs that has non-zero effect on the outcome. When C1 does not hold, it is subject to bias because the model ignores the missed set of $q - S$ PCs so that, now, the main effect of G is misspecified. Let $\boldsymbol{P}_i^s = (P_i^s, ..., P_i^s)^T$ be the stack of PCs corresponding to subject i. Result 4.3 explicitly gives the bias expression due to missing $q - S$ PCs.

Result 4.3 The bias due to fitting model (4.4) is given by

$$E_{H_0}[S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)] = \sum_{s=S+1}^{q} \{E[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{P}_i^s - \boldsymbol{\phi}^s\}\beta_{P,s}^0,$$

where $\beta_{P,s}^0$ is the coefficient in the full model where all PCs are included;

$$\begin{aligned}
\boldsymbol{\phi}^s &= E\{(\boldsymbol{E}_i * \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S]\}\boldsymbol{A}^{-1}\boldsymbol{b}^s \\
\boldsymbol{A} &= E\{[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S]^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S]\} \\
\boldsymbol{b}^s &= E\{[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S]^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{P}_i^s\}.
\end{aligned}$$

The result shows that the bias due to a PC that was not included is proportional to its association with the outcome conditional on $(\boldsymbol{X}_i, \boldsymbol{E}_i)$. This is also closely related to the definition of confounders discussed by VanderWeele and Shpitser (2013).

To reduce the bias due to the confounding effect of G, a better approach should consider the correlation between the outcome and the PCs in addition to the eigenvalues. A well-known method that takes this correlation into account is the partial least squares regression (PLS). PLS generates orthogonal components by sequentially optimizing their correlation with the outcome and correlation with G (Boulesteix and Strimmer (2007)). However, when the sample size is small and the region is large, PLS components are constructed by overfitting an outcome regression model, which makes the test for the interaction terms less powerful (Supplementary Table C.2). Instead, we propose to use the components $\{P_i^s\}_{1 \leq s \leq q}$ from PCA but rank them by

$$\mathrm{corr}(Y_{i,j}, P_i^s | \boldsymbol{X}_i, \boldsymbol{E}_i)^2 var(P_i^s) = R_s^2 \kappa_s,$$

where R_s^2 stands for the variation of $Y_{i,j}$ explained by P_i^s conditional on $(\boldsymbol{X}_i, \boldsymbol{E}_i)$. It is reasonable to assume R_s^2 is not likely to vary across visits j under model (4.1) because G is time invariant and we do not consider the situation that the association between G and Y may vary by visit j under the null hypothesis. This weighted PCA approach uses a criterion that is close to the objective function of PLS, but the selected PCA components are not constructed by fitting an outcome regression model. Similar approach of using correlation-selected PCs was also successfully used in GWAS to find PCs for population stratification adjustment (Lee et al. (2011)). To adjust for the effect of the exposure and covariates, we first regress $\boldsymbol{Y}_i$ on $(\boldsymbol{X}_i, \boldsymbol{E}_i)$, then use the residuals to estimate R^2 for each principal component. To reduce the dimension of the fitted model (4.4), we again suggest to use $S = m^{\frac{1}{2}}$ in practice to have reliable asymptotic estimation of $\boldsymbol{\beta}$ and illustrate it using extensive simulation studies (Wang (2011)).

4.5 Numerical Studies

We evaluated type I error rate and power of the proposed test using simulation studies for both cross-sectional and longitudinal data, and compared our method with existing choices: 1. set based tests for GEI using a single average or baseline outcome and exposure measure: iSKAT with $\rho = 0$ and rareGE assuming a random main effect of G (Lin et al. (2013); Lin et al. (2016); Chen et al. (2014)); 2. a single SNP based test for longitudinal outcomes and exposures: the minimum p-value test (MinP) using GEE. For each simulated dataset, we directly sampled SNPs from gene regions in MESA and then conditionally simulated the phenotype and environmental exposure. When there are repeated measurements, we first simulated the complete data, and then applied a missingness indicator with 4% fixed drop-out rate at each exam assuming data missing completely at random. The coefficients in the simulation studies were chosen such that each variable explains a reasonable variation in the outcome as in real data scenarios. For example, the variation in the outcome explained by the main effect of E or G (a set of SNPs) ranges from 5% to 15% in the longitudinal settings. We simulated top four principal components as covariates directly from MESA genome-wide data to retain its correlation with the target region, and their coefficients were elicited based on the analysis of the corresponding ethnic group. The simulation studies are structured into three scenarios where each part empirically evaluates both type I error and power based on 1000 replicates.

Scenario 1: Role of main effect specification of E. In the first simulation setting, we evaluated the proposed method when the main effect of E is linear/nonlinear in both cross-sectional and longitudinal settings. We focused on cubic-spline sieves generated by knots at equally spaced quantiles of all observations. We used all SNPs from region indexed by $rs10850411$ $(190$ SNPs) in European Americans $(2526$ subjects), and simulated one environmental exposure independent/dependent of the SNPs. To focus on the effect of E, this region was chosen such that the sample size is sufficiently large relative to the number of SNPs. The true model is of the form:

$$E_{i,j} = \alpha_{E,0} t_{i,j} + \alpha_{E,1} X_i + \sum_{k=1}^{5} \alpha_{E,2} G_i^k + b_{E,i} + \varepsilon_{E,i,j}, \quad j = 1, \ldots, d,$$

$$Y_{i,j} = \sum_{s=1}^{4} \alpha_{PC,s} PC_i^s + \alpha_0 t_{i,j} + \alpha_1 X_i + \alpha_2 h_M(E_{i,j}) + \sum_{k=1}^{5} \alpha_3 G_i^k + \alpha_4 E_{i,j} \sum_{k=1}^{5} G_i^k + b_i + \varepsilon_{i,j},$$

where $d = 1$ is for cross-sectional data and $d = 4$ is for longitudinal data; $t_{i,j} = j - 1$ (0, 1, 2, 3 standing for visits); $X_i \sim N(0, 1)$ is a time-invariant covariate; PC_i^s is the s-th principal component of subject i directly from the MESA genome-wide data; five out of the 190 SNPs (2.6%) are causal and G_i^k is the genotype of subject i for the k-th randomly selected causal SNP; $(\alpha_{PC,1}, \alpha_{PC,2}, \alpha_{PC,3}, \alpha_{PC,4}) = (-4.7, -0.9, 13.1, 1.3)$; $\alpha_{E,0} = \alpha_{E,1} = \alpha_0 = \alpha_1 = 1$, $\alpha_2 = 0.5$, $\alpha_3 = 2$; $\alpha_{E,2}$ measure the association between E and G. $\alpha_{E,2} = 0$ when E is independent of G and $\alpha_{E,2} = 0.5$ when E is dependent of G (e.g., $\sim 3\%$ variation in E is explained by G in the longitudinal setting); $\alpha_4 = 0.10/0.05$ for evaluating cross-sectional/longitudinal power and $\alpha_4 = 0$ for evaluating type I error rate; $b_{E,i} \sim N(0, 4)$, $\varepsilon_{E,i,j} \sim N(0, 4)$, $b_i \sim N(0, 9)$, $\varepsilon_{i,j} \sim N(0, 9)$ and they are all independent. h_M is the main effect function specified as "E", "$0.3E^2$", "$E + 0.2E^2$" or "$\exp(0.4E)$" for cross-sectional data, and "$0.8E$", "$0.2E^2$", "$0.5E + 0.1E^2$" or "$\exp(0.3E)$" for longitudinal data. The functions were scaled such that they explain similar variation of $Y_{i,l}$ as compared to the linear model (e.g., $\sim 10\%$ in the longitudinal setting). Table 4.1 presents the results.

Type I error rate. Even when C1 holds, iSKAT using a model based inference has inflated type I error rate (e.g., 0.172, 0.113 and 0.185 where the true models are E^2, $E + E^2$ and $\exp(E)$ respectively, cross-sectional setting). rareGE has inflated type I error rate when the main effect of E is nonlinear, similar to iSKAT. However, the proposed method using the sandwich estimator is robust regardless of the main effect misspecification; When C1 does not hold, only assuming a linear main effect does have type I error inflation even if sandwich estimation is used (e.g., 0.906, 0.729 and 0.869 for E^2, $E + E^2$ and $\exp(E)$, longitudinal setting). However, the proposed method using the method of sieves still has robust type I error rate.

Power. When C1 holds, the proposed method using the method of sieves always has similar power as the method based on the true model, even if the true effect is linear and additional model complexity was assumed for the main effects (e.g., 0.786 vs. 0.789, longitudinal setting). When C1 is violated, the method of sieves results in slightly lower power than using the true model (e.g., 0.774 vs. 0.796, longitudinal setting), but the power difference is small. Moreover, the method of sieves often leads to improved power compared with the method assuming a linear main effect when the true effect is nonlinear (e.g., 0.786 vs. 0.606 when the true main effect is E^2, longitudinal setting).

Scenario 2: Role of main effect specification of G. In the second simulation setting, we evaluated the proposed method for the main effect adjustment of G in both cross-sectional and longitudinal settings. We varied the number of SNPs (400 - 700) simulated from genotype region $MECOM$ (821 SNPs) in Chinese Americans (775 subjects), and simulated one environmental exposure independent/dependent of the SNPs. The region was chosen to reflect a scenario where the number of SNPs is large relative to the sample size. The model is same as that in Scenario 1 with a linear main effect of E, so we omit the detailed equations and only present the parameters that are different from Scenario 1. In this scenario, five out of the 400/700 SNPs (1.3%/0.7%) are causal; $(\alpha_{PC,1}, \alpha_{PC,2}, \alpha_{PC,3}, \alpha_{PC,4}) = (-2.3, -24.9, 5.6, -13.3)$; $\alpha_{E,2} = 0$ when E is independent of G and $\alpha_{E,2} = 2$ when E is dependent of G (e.g., $\sim 25\%$ variation in E is explained by G in the longitudinal setting). We chose a large $\alpha_{E,2}$ to observe the type

I error inflation due to main effect misspecification of G. $\alpha_4 = 0.2/0.1$ for evaluating cross-sectional/longitudinal power and $\alpha_4 = 0$ for evaluating type I error rate. The results are summarized in Table 4.2.

Type I error rate. The MinP test based on single SNP analyses has inflated type I error rate when C1 does not hold (e.g., 0.088/0.108 for 400/700 SNPs, longitudinal setting). This result is consistent with the results in Lin et al. (2013). iSKAT using a ridge regression has type I error inflation when the number of SNPs is large, especially when the number is close to the sample size (0.089 for 700 SNPs, cross-sectional setting). Supplementary Table C.1 further shows an example where iSKAT has type I error rate close to one when the number of SNPs exceed the sample size. rareGE has slightly inflated type I error rate when the number of SNPs is greater than the sample size (0.072 for 700 SNPs, cross-sectional setting, Supplementary Table C.1). The proposed method has well controlled type I error rate for all scenarios considered in this stimulation setting. We further evaluated PCA, PLS and weighted PCA as other possible approaches to reduce dimension of G and summarized the results in Supplementary Table C.2. When C1 holds and the number of adjusted components is five, type I error rates of PLS and weighted PCA are well controlled, but that of PCA is inflated (e.g., 0.033/0.057/0.094 for PLS/weighted PCA/PCA, 700 SNPs, longitudinal setting). When the number of components increases to $m^{\frac{1}{2}}$, all three have well controlled type I error rate. The proposed methods tend to be slightly conservative due to the use of sandwich estimator as in regular GEE, even if a correct mean model is used.

Power. The proposed method has similar power as using the true model and it is more powerful than the MinP test (e.g., 0.588 vs. 0.472, 400 SNPs, longitudinal setting when E and G are independent). We also evaluated the proposed test using a model based inference for estimating $\boldsymbol{\Sigma}$ that is typically used for cross-sectional data. It has slightly higher power than using the sandwich estimation (e.g., 0.626 vs. 0.588, 400 SNPs, longitudinal setting when C1 holds). The power of rareGE is comparable to our proposed test using a model based inference in situations when there are no Type 1 error inflation (for example, in Table 2, both are equal (0.561) when C1 holds and the number of SNPs is 700). Moreover, Supplementary Table C.2 shows that PLS has lower power than the proposed method when the number of SNPs is close to the sample size (e.g., 0.381 vs. 0.483, 700 SNPs, cross-sectional setting when C1 holds), although its type I error rate is well controlled.

Scenario 3: Role of longitudinal data. In the third simulation setting, we aimed to illustrate that the proposed method is robust to misspecification of within-subject correlation when there are repeated measurements, and show the advantage of using full trajectory of the longitudinal outcome and exposure. When more than one repeated measures are involved, we compare our method with iSKAT using the average/baseline value of the repeated measurements on both Y and E. We used all SNPs from genotype region indexed by $rs10850411$ (190 SNPs) in European Americans (2526 subjects), and simulated one environmental exposure independent of the SNPs. The model is same as the longitudinal setting in Scenario 1 with an linear main effect of E, so we omit the detailed equations and only present the parameters that are different from Scenario 1. In this scenario, $\alpha_4 = 0.05$ for evaluating power and $\alpha_4 = 0$ for evaluating type I error rate; $b_{E,i} \sim N(0, 0.25)$, $\varepsilon_{E,i,j} \sim N(0, 4)$, $b_i \sim N(0, 9/16)$, $\varepsilon_{i,j} \sim N(0, 9)$ and they are all independent. We note that we simulated a large magnitude of within-subject variation to show the type I error inflation due to using the average value of repeated measures. The relative power difference remains the same when a smaller within-subject variation is simulated. Table 3 presents the results.

Type I error rate. The proposed method using the first order autoregressive correlation structure still has valid type I error rate, when the true correlation structure is compound symmetric. iSKAT and rareGE using the average value of repeated measurements has inflated type I error rate because of their model based inference and the heterogeneous variance due to unbalanced data structure (e.g., 0.092 for iSKAT and 0.078 for rareGE, $d = 4$).

Power. The tests using the full trajectory of longitudinal outcome and exposure have much higher power than using the average values, as the number of repeated measurements increases (e.g., 0.805 vs. 0.414, $d = 4$). This is because averaging the environmental exposure reduces its variance and therefore decreases the power of testing gene-environment interaction. The results demonstrate the advantage of using the longitudinal information.

Table 4.1: Simulation study evaluating the main effect adjustment of E (2526 subjects, 190 SNPs). iSKAT: set based test proposed by Lin et al. (2013). rareGE: rareGE test proposed by Chen et al. (2014) assuming a random main effect of G. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using natural cubic-spline smoothing for E with $\sqrt{m}$ basis functions. GE-true: the proposed test with the correct model, which correctly specifies the main effect of E. The GE methods were implemented using the weighted PCA approach for the main effect of G. Each cell presents type I error rate or power based on 1000 replicates evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07. The calibrated powers with value zero correspond to very high type I errors.

Cross-sectional data								
Type I error rate								
	C1 holds				C1 does not hold			
$h_M(E)$	E	E^2	$E+E^2$	$\exp(E)$	E	E^2	$E+E^2$	$\exp(E)$
iSKAT	0.055	0.170	0.109	0.181	0.054	0.899	0.751	0.947
rareGE	0.057	0.174	0.108	0.195	0.050	0.905	0.762	0.951
GE-linear	0.055	0.046	0.046	0.032	0.050	0.780	0.663	0.618
GE-spline	0.053	0.053	0.054	0.057	0.050	0.051	0.048	0.051
GE-true	0.053	0.053	0.052	0.055	0.051	0.050	0.048	0.053
Power								
	C1 holds				C1 does not hold			
$h_M(E)$	E	E^2	$E+E^2$	$\exp(E)$	E	E^2	$E+E^2$	$\exp(E)$
iSKAT	0.751	0.549*	0.666*	0.553*	0.761	0*	0*	0*
rareGE	0.760	0.565*	0.681*	0.553*	0.771	0.329*	0.445*	0.185*
GE-linear	0.746	0.556	0.669	0.572	0.754	0.949	0.932	0.829
GE-spline	0.745	0.745	0.750	0.741	0.733	0.726	0.727	0.721
GE-true	0.750	0.747	0.754	0.743	0.756	0.740	0.739	0.744
Longitudinal data ($d = 4$)								
Type I error rate								
	C1 holds				C1 does not hold			
$h_M(E)$	E	E^2	$E+E^2$	$\exp(E)$	E	E^2	$E+E^2$	$\exp(E)$
GE-linear	0.045	0.042	0.040	0.039	0.049	0.906	0.729	0.869
GE-spline	0.043	0.042	0.043	0.042	0.048	0.048	0.048	0.048
Power								
	C1 holds				C1 does not hold			
$h_M(E)$	E	E^2	$E+E^2$	$\exp(E)$	E	E^2	$E+E^2$	$\exp(E)$
GE-linear	0.793	0.606	0.731	0.681	0.796	0*	0*	0*
GE-spline	0.786	0.786	0.786	0.784	0.774	0.774	0.775	0.769
GE-true	0.789	0.788	0.788	0.789	0.796	0.780	0.780	0.782

Table 4.2: Simulation study evaluating the main effect adjustment of G (775 subjects). A linear main effect of E was fitted. Each cell presents the type I error rate/power based on 1000 replicates. MinP: single SNP analysis using GEE adjusted by the effective number of independent tests (Gao et al., 2008). iSKAT: region based test proposed by Lin et al. (2013). rareGE: rareGE test proposed by Chen et al. (2014) assuming a random main effect of G. GE-none: the proposed test adjusting for none of the SNPs. GE-wPCA/wPCAM-$\sqrt{m}$: the proposed test adjusting for the leading $\sqrt{m}$ components using the weighted PCA and robust(wPCA)/model-based(wPCAM) inference. GE-true: the proposed test with the correct model, which correctly correctly includes all SNPs with non-zero main effects. Type I error rate and power were both evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07.

Cross-sectional data								
	Type I error rate				Power			
	C1 holds		C1 does not hold		C1 holds		C1 does not hold	
q	400	700	400	700	400	700	400	700
MinP	0.052	0.047	0.116	0.129	0.426	0.355	0.676*	0.636*
iSKAT	0.050	0.066	0.055	0.060	0.557	0.505	0.780	0.747
rareGE	0.054	0.055	0.072	0.059	0.634	0.561	0.821*	0.810
GE-none	0.025	0.038	0.189	0.181	0.446	0.402	0.769	0.778
GE-wPCA -$\sqrt{m}$	0.030	0.038	0.040	0.030	0.540	0.514	0.771	0.741
GE-wPCAM-$\sqrt{m}$	0.041	0.038	0.048	0.048	0.591	0.561	0.805	0.785
GE-true	0.036	0.038	0.045	0.041	0.584	0.509	0.797	0.759

Longitudinal data								
	Type I error rate				Power			
	C1 holds		C1 does not hold		C1 holds		C1 does not hold	
q	400	700	400	700	400	700	400	700
MinP	0.028	0.033	0.088	0.108	0.472	0.462	0.568*	0.575*
GE-none	0.045	0.039	0.187	0.173	0.560	0.564	0.477*	0.435*
GE-wPCA -$\sqrt{m}$	0.033	0.034	0.031	0.034	0.588	0.569	0.687	0.682
GE-wPCAM-$\sqrt{m}$	0.034	0.037	0.032	0.040	0.626	0.589	0.736	0.708
GE-true	0.040	0.034	0.038	0.039	0.615	0.589	0.724	0.707

Table 4.3: Simulation study evaluating the use of longitudinal data. GE-CS/AR1: the proposed test with different working correlation (compound symmetric/first order autoregressive). GE-avg.: the proposed test using the average/baseline value of repeated measurements. iSKAT-avg.: cross-sectional iSKAT using the average value of repeated measurements. rareGE-avg.: rareGE test proposed by Chen et al. (2014) using the average value of repeated measurements, and assuming a random main effect of G. The GE methods were implemented using the weighted PCA approach for the main effect of G and natural cubic-spline for the main effect of E adjusting for $\sqrt{m}$ terms. Each cell presents type I error rate or power based on 1000 replicates evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07.

Type I error rate					
	Methods using full trajectory		Methods using average value		
d	GE-CS	GE-AR1	GE-avg.	iSKAT-avg.	rareGE-avg.
1	0.047	0.047	0.047	0.047	0.048
2	0.046	0.046	0.041	0.050	0.050
3	0.048	0.048	0.058	0.072	0.063
4	0.039	0.041	0.052	0.092	0.078

Power					
	Methods using full trajectory		Methods using average value		
d	GE-CS	GE-AR1	GE-avg.	iSKAT-avg.	rareGE-avg.
1	0.379	0.379	0.379	0.387	0.400
2	0.581	0.581	0.390	0.415	0.430
3	0.722	0.719	0.431	0.422*	0.466
4	0.805	0.803	0.414	0.409*	0.456*

4.6 Data Analysis

We illustrate the proposed set-based test using data from the Multi-Ethnic Study of Atherosclerosis (MESA) to test the interaction between each neighborhood variable and each SNP set on blood pressure (systolic and diastolic blood pressure) for the four ethnic groups separately, followed by a meta-analysis. Supplementary Tables C.3 - C.8 present the summary statistics and the marginal association analysis of E/G in MESA. Density of favorable food stores, density of recreational facilities and perceived healthy foods availability are marginally significantly associated with systolic blood pressure (p-value = 3.65×10^{-3}, 4.69×10^{-4} and 8.74×10^{-7} respectively); Perceived healthy foods availability is also associated with diastolic blood pressure (p-value = 4.21×10^{-3}). The marginal effects of the environmental exposures appear to be mostly linear. We conducted both a targeted GEI analysis for the 29 candidate regions and a set-based genome-wide GEI analysis as described in section 4.2. We adjusted for age, gender, body mass index (BMI), a socioeconomic status variable (SES) and top four ethnicity-specific principal components (PCs) to correct for potential within-ethnicity stratification. BMI was calculated from direct measurements of weight (kg) and height (meters) available for all MESA exams. The socioeconomic status variable was obtained by performing a principal component analysis on a set of housing, residential stability, education, employment, occupation and income variables. We adjusted for the first leading component which is more highly weighted on education, occupation and income. We adjusted the measured blood pressures for participants taking anti-hypertension medication using the standard procedure of adding 10 mmHg to systolic blood pressure and 5 mmHg to diastolic blood pressure as in Cui et al. (2003). Based on the p-values of the ethnicity-stratified analysis, a meta-analysis was done by Fisher's combined probability test (Fisher (1925)). The vast majority of SNPs in our dataset are common variants with MAF greater than 1%, therefore the genetic principal components calculated for each region in the weighted PCA approach mostly capture the genetic variation in common variants.

Targeted GEI analysis. We conducted a set-based analysis for the 29 candidate genomic regions and compared our method with GEE-based MinP test and iSKAT using either the average or baseline value of repeated measurements. This set-based analysis led to 29 sets $\times$ 4 exposures $= 116$ tests. We also conducted a single SNP analysis for all SNPs in the 29 regions that led to 5622 SNPs $\times$ 4 exposures $= 22488$ tests and present the results using the locus-zoom plots (Supplementary Figure C.2) (Pruim et al. (2010)).

Set-based analysis. Table 4.4 presents the most significant region identified by the set-based analysis. The proposed methods exhibit highly suggestive p-value (0.0005 using a linear main effect of E, 0.0009 using the natural cubic-spline) for the interaction between perceived healthy food availability and the region indexed by rs10850411 on systolic blood pressure in European Americans. These p-values are very close to the Bonferroni threshold $(0.05/(4 \times 29) = 0.00043)$. MinP test also results in a suggestive p-value (0.0047) but iSKAT and rareGE using average/baseline value fail to identify this interaction (p-value = 0.8205 and 0.4331 for iSKAT, 0.7542 and 0.5336 for rareGE respectively). This interaction is also suggestive for its GIS counterpart (density of favorable food stores) by using the proposed method (p-value = 0.0427 using a linear main effect of E, 0.0570 using the natural cubic-spline). The most significant SNP in this region is the index SNP rs10850411 (p-value = 4.08×10^{-5}). The locus-zoom plot (the left panel in Supplementary Figure C.2) shows there are multiple other SNPs with small p-values uniformly distributed in the region and they are in linkage disequilibrium with the index SNP. We conducted sensitivity analysis additionally adjusting for site and present the results for this region in Supplementary Table C.9. The results with and without adjusting for study site are qualitatively similar with some small numerical differences. We also conducted additional analyses to compare strategies using different forms of longitudinal exposures and present results in Supplementary Table C.10. The results show that using repeated longitudinal measures appear to be a better strategy in general.

Single-SNP analysis. The most significant SNP identified by the single-SNP analysis is the interaction between density of recreational facilities and a SNP in region $CACNB2$, namely rs7085587, on systolic blood pressure in Hispanic Americans (p-value = 2.17×10^{-6}). This p-value is still significant after the Bonferroni correction (Bonferroni threshold: $0.05/22488 = 2.22 \times 10^{-6}$). The locus-zoom plot (the right panel in Supplementary Figure C.2) shows the signals are concentrated in a small area around rs7085587. This is also a situation where the MinP test results in a smaller p-value (0.0012) than the proposed test (GE-linear p-value = 0.0753) in the corresponding set-based analysis of $CACNB2$ (Supplementary Table C.11).

In summary, the set-based test performs better in the first example where the signals are dispersed across many SNPs in the region, while the single SNP based test performs better in the second example where the signals are concentrated. For the significant interactions noted with systolic blood pressure as outcome in Table 4, we also observed suggestive p-values (0.0844 using a linear main effect of E, 0.0537 using the natural cubic-spline for the main effect of E) for diastolic blood pressure (Supplementary Table C.12). These interactions that appear to be noteworthy in the European Americans in the MESA analysis are ethnicity specific and are not significant in the other three ethnic groups. The present finding will require replication in other cohorts and need to be followed up in future studies. The genes nearest to rs10850411 are two members of the phylogenetically conserved T-box family of genes, $TBX3$ and $TBX5$. Proteins encoded by T-box family genes act as transcription factors, and have been shown to play a role in development of the heart and limbs (McKusick (1998); OMIM 601620, 601621). Genome-wide association studies have identified variants in the $TBX3$-$TBX5$ gene region that influence heart rate and cardiac electrical activity (Pfeufer et al. (2010); Sotoodehnia et al. (2010)). $CACNB2$ encodes the beta-2 subunit of a voltage-dependent calcium channel protein, and is expressed in the heart. Mutations in $CACNB2$ have been shown to cause Brugada syndrome, characterized by cardiac electrical abnormalities and sudden cardiac death (OMIM 600003, 611876). The detailed analysis of the top SNPs in these two identified regions can be found in section 4.7 of the Supplementary Materials (Appendix C, Supplementary Figure C.3).

Genome-wide GEI analysis. We applied the proposed test to 24743 genes (Section 2) for a set-based analysis and compared it with a single SNP analysis of 1011876 SNPs in these sets via GEE. Supplementary Figures C.4 - C.5 presents the QQ-plots summarizing the results of the set-based analysis using our proposed method. The set based analysis identified a highly suggestive interaction between region $LOC100129138$ and perceived walkability on systolic blood pressure (p-value = 2.04×10^{-6}, Bonferroni threshold = 2.02×10^{-6}). However, the single SNP analysis did not identify any interaction between any SNP and perceived walkability. The smallest p-value equals 8.12×10^{-6} which is much higher than the Bonferroni threshold $= 4.94 \times 10^{-8}$. This illustrates the potential advantage of a genome-wide set-based GEI analysis compared to a genome-wide single SNP-based GEI analysis. In addition, we observed that iSKAT QQ plots are substantially inflated for a genome-wide analysis in MESA (Supplementary Figures C.6 - C.7), which is consistent with our simulation studies. This is because the CHN ethnic group only has 775 subjects, but there are many large regions in the genome-wide analysis. The performance of iSKAT with $m < q$ is less than optimal.

Table 4.4: Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data: interactions between neighborhood variables and the region indexed by rs10850411 on systolic blood pressure. Each cell shows the p-value. EUR: European Americans; AFA: African Americans; HIS: Hispanics; CHN: Asians of Chinese descent. Meta: Meta-analysis combining the results of four ethnic groups using Fisher's combined probability test. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E. MinP: minimum p-value test based on GEE. The assumed working correlation is compound symmetric. iSKAT-avg./base.: cross-sectional iSKAT using the average/baseline value of repeated measurements as the outcome. rareGE-avg./base.: cross-sectional rareGE using the average/baseline value of repeated measurements as the outcome, where a random main effect of G is assumed. Bonferroni correction threshold is 0.00043.

Systolic Blood Pressure - Region Indexed by rs10850411 (190 -214 SNPs)										
	Density of favorable food stores					Density of recreational facilities				
	EUR	CHN	AFA	HIS	Meta	EUR	CHN	AFA	HIS	Meta
GE-linear	0.0427	0.0890	0.8651	0.6256	0.1353	0.7857	0.9631	0.4937	0.8130	0.9670
GE-spline	0.0570	0.0544	0.9320	0.7231	0.1366	0.8480	0.9640	0.5405	0.8891	0.9848
MinP	0.0602	0.5753	1.0000	1.0000	0.5664	1.0000	1.0000	1.0000	1.0000	1.0000
iSKAT-avg.	0.2416	0.3695	0.7435	0.9257	0.6942	0.5134	0.3400	0.3869	0.5226	0.5707
iSKAT-base.	0.3953	0.2459	0.7200	0.9298	0.7070	0.7215	0.3239	0.5886	0.7561	0.8068
rareGE-avg.	0.2421	0.3144	0.9448	0.7851	0.6754	0.5281	0.5386	0.5169	0.3989	0.6839
rareGE-base.	0.4524	0.1045	0.9670	0.8334	0.5875	0.8591	0.3717	0.5458	0.7175	0.8426
	Perceived Healthy Food Availability					Perceived walkability				
	EUR	CHN	AFA	HIS	Meta	EUR	CHN	AFA	HIS	Meta
GE-linear	0.0005	0.9736	0.7591	0.9270	0.0446	0.2812	0.3235	0.3384	0.1678	0.2297
GE-spline	0.0009	0.9067	0.8241	0.9034	0.0608	0.2127	0.3746	0.3166	0.2058	0.2303
MinP	0.0047	1.0000	1.0000	1.0000	0.2177	1.0000	1.0000	1.0000	1.0000	1.0000
iSKAT-avg.	0.8205	0.4296	0.4285	0.6091	0.7817	0.9028	0.6702	0.6318	0.7490	0.9617
iSKAT-base.	0.4331	0.5049	0.4503	0.5283	0.6571	0.8422	0.8531	0.6441	0.6475	0.9658
rareGE-avg.	0.7542	0.3500	0.5124	0.2821	0.5878	0.8363	0.4632	0.6999	0.6264	0.8956
rareGE-base.	0.5336	0.2642	0.3120	0.2036	0.3073	0.9141	0.7050	0.6569	0.3873	0.8900

CHAPTER V

Summary and Discussion

Chapter II presents a novel framework for modeling and testing for the joint association of genetic variants with a trait from the perspective of viewing traits as a random field on a genetic space. The development has been focused on quantitative traits with a normal distribution. Based on the GenRF model, a test for genetic associations was developed and this test enjoys many appealing features. The GenRF test is based on testing a null hypothesis involving a single parameter, allowing it to exploit LD to improve power. When LD is moderate or high, our simulations showed that the GenRF test achieves much higher power than the more traditional regression-based methods. The GenRF model is flexible to allow for complex interaction effects and, as demonstrated by simulations, the GenRF test is even much more powerful than SKAT in the presence of complex interaction effects. Moreover, as SKAT, prespecified variant-specific weights can be incorporated to boost power for rare variants. Unlike SKAT, the GenRF test is an exact test under the normal assumption and thus not overly conservative in finite samples. Finally, the test is computationally easy to implement since an analytical form is available. In summary, the GenRF test is an appealing alternative to SKAT and other existing methods for testing the joint association of variants with a trait. It can achieve overall comparable performance and sometimes even much better performance relative to SKAT as well as other methods.

Although we focus on quantitative traits, we note that the GenRF test is robust to distributions other than normal as discussed previously and demonstrated by simulation studies. Specifically for binary traits, although the GenRF model with an identity link function may seem a bit unnatural, the resulting test with a minor modification is still valid and can achieve good power. However, due to the conceptual difficulty associated with modeling binary traits using a linear model and the possible practical issue that can arise, it would be interesting to study, within the framework of random field model, other link functions for binary traits as well as other distributions in the future.

Chapter III extended the genetic random field model to the longitudinal setting and developed generalized score type tests to test the joint association between a set of genetic variants and a repeatedly measured phenotype. Besides the advantages of region-based tests over single-marker tests in cross-sectional studies, the LGRF model is able to utilize all the repeated measurements, incorporate gene-time interaction explicitly and result in higher power. As in GenRF, LGRF models the joint association using a single parameter by considering the similarity in phenotype induced by genetic similarity. A main challenge in modeling longitudinal data is to account for within-subject correlation and correlation is conceptually viewed and modeled in a unified way as the joint genetic association in LGRF. Furthermore, the specified correlation structure is treated as a working assumption in inference and the resulting LGRF tests are robust to misspecification.

LGRF tests are generalized score tests that only need to fit the model under the null hypothesis, which is irrelevant to the target region. Users can fit the null model once and test all regions without repeatedly fitting the model. In addition, the computational cost of LGRF mainly depends on the fixed number of variants in the region but not the sample size. This property improves the computational efficiency dramatically (see Supplementary Table B.1) especially when the target region is small, for example if investigators are only interest in the exon.

Chapter IV focuses on set-based inference for testing gene-environment interaction with quantitative traits in both cross-sectional and longitudinal studies. We showed that a generalized score test similar to the tests derived from more sophisticated approaches (e.g., kernel machine regression, similarity regression and genetic random field model) could be postulated using the most commonly used fixed effect model for multivariate regression. Instead of a hybrid model like iSKAT where the main effects are considered as fixed effects but the interactions are considered as random effects, the proposed fixed effect model presents a direct unified framework. We also demonstrated improved properties of a set-based test compared to a single SNP analysis when multiple causal SNPs exist.

Although many set-based tests have been proposed for evaluating genetic association, our test is the first set-based test for GEI that is able to utilize the rich time varying outcome and exposure data. Our numerical studies show that substantial power gain can be achieved by using the proposed test, compared to methods only using a single outcome/exposure measure (e.g., average/baseline value). The test is also robust to misspecification of within-subject correlation, which is a desirable property in studies with longitudinal measures.

We studied the role of gene-environment independence, and developed methods for main effect adjustment of E and G that permits more robust and powerful inference. Under the independence condition, we showed that the proposed test is robust to misspecification of main effect of E by simply using a sandwich estimator and weighted centered E and G. When the independence condition does not hold, we proposed the method of sieves to model the main effect of E correctly. An interesting finding is that flexibly modeling the main effect does not hurt power for tests of GEI significantly. To remedy the curse of dimensionality in the potentially high dimensional G space, we developed the weighted PCA approach for dimension reduction that allows us to apply the test to large regions where the number of SNPs is close to or larger than the sample size.

We illustrated the method by a targeted GEI analysis and a genome-wide GEI analysis of MESA neighborhood study, where both time varying outcome and exposure data are available. The application illustrates that the longitudinal approach utilizing the full trajectory of longitudinal outcome and exposure measures is substantially more powerful than the approach using a single measurement. It also shows the advantage of a genome-wide set-based GEI analysis compared to a genome-wide single SNP-based GEI analysis. The application is novel in its rich longitudinal neighborhood data and the findings may aid in prevention of chronic diseases by modifying the built environment around us and creating new healthy food resources and recreational facilities and provide public health recommendations for susceptible genetic sub-groups in terms of their neighborhood choice. More importantly, neighborhood interventions or changes in the built environment can impact many people at the same time instead of recommending changes towards lifestyle factors of an individual.

There are several limitations of the proposed method. First, the weighted PCA method is an ad-hoc method proposed for dimension reduction of G, only studied through simulation and data analyses. The optimality of this method has not been established in this paper. It will be desirable to develop an optimal method for the main effect adjustment of G in the future and establish its theoretical properties more rigorously. Second, we only considered linear GEI terms in this paper. Directly adding more flexible non-linear GEI terms will certainly lead to loss of power, which is different from flexibly modeling the main effects. It will be interesting to investigate efficient strategies for modeling non-linear interaction terms. Third, the method was proposed for quantitative traits. Future extension to generalized linear models will be important to develop. Moreover, we note that our result is closely related to the work by Vansteelandt et al. (2008), where they proposed multiply robust inference for statistical interactions by not only modeling the main effect of E, but also the conditional distribution of E given X and G. Future work that develops a multiply robust set-based inference for GEI boosted with dimension reduction in the G space will be of great interest.

APPENDICES

APPENDIX A

Supplementary Materials for Chapter II

1. Robustness to other distributions

We evaluated the robustness of the GenRF test to distributions other than normal. The GenRF test for traits with distributions other than normal is described in Section 2.3 of the main manuscript. The simulation setup is otherwise similar to the first set of simulations, described in Section 3 of the main manuscript, with only one region, $p = 10$, $\rho = 0.4$ and $n = 100$. Responses Y_i were generated according to generalized linear models using the canonical link function, i.e.,

$$g(\mu_i) = aG_{i,5},$$

where a was set to be 1.1 and 2.5 respectively for exponential and binary distributions. For Mixture Normal, we generated two normal distributions with mean difference 10, equal mixture proportions, and $a = 2.7$. We set a to be 0 in evaluating the type-I error rate. The results are shown in Supplementary Table A.1.

Table A.1: Simulation results under different distributions of the response variable (1000 replicates). $*$ indicates results are unavailable due to "sample size is small, need small sample adjustment" and SKAT has no small sample adjustment for IBS kernel.

Method		Distribution		
		Exponential	Mixture Normal	Binary
GenRF	Power	0.636	0.582	0.646
	Type I	0.052	0.056	0.046
SKAT	Power	0.655	0.582	$*$
	Type I	0.046	0.046	$*$
F-test	Power	0.572	0.568	0.559
	Type I	0.056	0.054	0.050

2. Robustness to heteroscedastic variances of binary traits

We evaluated the robustness of the GenRF test to heteroscedasitc variances of binary traits. Since the variance of a binary outcome is a function of its mean, the variance is known to be heteroscedastic when the mean of outcome depends on covariates. The modification of the GenRF test for binary traits is described in Section 2.3 of the main manuscript. The simulation setup is otherwise similar to the first set of simulations, described in Section 3 of the main manuscript, with $p = 20$, $\rho = 0$, minor allele frequency 0.2, and $n = 100$. Responses Y_i were generated according to logistic models, i.e.,

$$\text{logit}(p_i) = aG_{i,5} + bX_i,$$

where a was set to be 3 in evaluating power and 0 in evaluating type I error rate; X_i was a covariate generated from $N(0, 1)$; and b was varying from 0 to 10 to generate different levels of heteroscedastic variance. A larger coefficient b results in a wider range of the predicted mean and thus more heteroscedastic variance. When $b = 5$ or 10, which represents unusually strong effect of X (probably unlikely in practice), some predicted means fall outside of $[0, 1]$ and truncation at 0 or 1 was used. The results are shown in Supplementary Table A.2. We note that the power decreases as the coefficient b increases because the noise becomes larger. The type I error is well controlled even if some predicted means reached 0 or 1, indicating that the GenRF test with the minor modification is robust to heteroscedastic variances of binary traits.

Table A.2: Simulation results under different levels of heteroscedastic variances (500 replicates). Coefficient: the coefficient of the covariate. $*$ indicates that some predicted means reached 0 or 1.

Method		Coefficient				
		0	1	3	5$*$	10$*$
GenRF	Power	0.538	0.594	0.408	0.290	0.110
	Type I	0.052	0.046	0.040	0.058	0.060

3. Application to Dallas Heart Study

We analyzed data from the Dallas Heart Study (Browning et al., 2004.), a population-based, multi-ethnic study on 3551 subjects whose lipids and glucose metabolism are measured. In this study, sequence variations in the coding regions of the four genes, ANGPTL3, ANGPTL4, ANGPTL5 and ANGPTL6 are discovered. Supplementary Table A.3 lists the number of non-synonymous variants in each gene and their MAFs.

Table A.3: Dallas Heart Study sequencing data information: number of non-synonymous variants in each gene. MAF: minor allele frequency.

	Number of Variants			
	ANGPTL3	ANGPTL4	ANGPTL5	ANGPTL6
All	21	25	18	25
MAF $<$ 5%	21	24	18	25
MAF $<$ 1%	20	23	17	24

APPENDIX B

Supplementary Materials for Chapter III

1. Detailed Proofs

1.1 Unbiasedness of the Estimating Equations

We show that the estimating function

$$U_\gamma(\boldsymbol{\beta}, \eta, \gamma) = \frac{\partial E(\boldsymbol{Y}|\boldsymbol{Y}_-)}{\partial \gamma}^T \{\boldsymbol{Y} - E(\boldsymbol{Y}|\boldsymbol{Y}_-)\} = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{S}(\boldsymbol{I} - \eta\boldsymbol{T} - \gamma\boldsymbol{S})(\boldsymbol{Y} - \boldsymbol{\mu})$$

is unbiased and the generalized score evaluated at $\gamma = 0$, $U_\gamma(\boldsymbol{\beta}, \eta, 0) = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{S}(\boldsymbol{I} - \eta\boldsymbol{T})(\boldsymbol{Y} - \boldsymbol{\mu})$ has mean 0 under H_0 and positive mean $\gamma E\{(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{S}^2 (\boldsymbol{Y} - \boldsymbol{\mu})\}$ under $H_1 : \gamma > 0$. Below we denote $U_\gamma(\boldsymbol{\beta}, \eta, \gamma) = \sum_{i,l} U_{\gamma,i,l}(\boldsymbol{\beta}, \eta, \gamma)$ where

$$U_{\gamma,i,l}(\boldsymbol{\beta}, \eta, \gamma) = \frac{\partial E(Y_{i,l}|\boldsymbol{Y}_{-(i,l)})}{\partial \gamma} \{Y_{i,l} - E(Y_{i,l}|\boldsymbol{Y}_{-(i,l)})\}.$$

Using an iterated expectation argument, we have

$$\begin{aligned} E\{U_{\gamma,i,l}(\boldsymbol{\beta}, \eta, \gamma)\} &= E[E\{U_{\gamma,i,l}(\boldsymbol{\beta}, \eta, \gamma)\}|\boldsymbol{Y}_{-(i,l)}] \\ &= E[\frac{\partial E(Y_{i,l}|\boldsymbol{Y}_{-(i,l)})}{\partial \gamma} \{E(Y_{i,l}|\boldsymbol{Y}_{-(i,l)}) - E(Y_{i,l}|\boldsymbol{Y}_{-(i,l)})\}] = 0, \end{aligned}$$

where the second equality is because $\frac{\partial E(Y_{i,l}|\boldsymbol{Y}_{-(i,l)})}{\partial \gamma}$ is a function of $\boldsymbol{Y}_{-(i,l)}$. Therefore, under correct specification of the model, i.e., $E(\boldsymbol{Y}|\boldsymbol{Y}_-) = \boldsymbol{\mu} + (\eta\boldsymbol{T} + \gamma\boldsymbol{S})(\boldsymbol{Y} - \boldsymbol{\mu})$, the estimating function $U_\gamma(\boldsymbol{\beta}, \eta, \gamma) = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{S}(\boldsymbol{I} - \eta\boldsymbol{T} - \gamma\boldsymbol{S})(\boldsymbol{Y} - \boldsymbol{\mu})$ is unbiased in the sense that it has expectation zero. Because of this unbiasedness, it follows that

$$E\{U_\gamma(\boldsymbol{\beta},\eta,0)\} = E\{U_\gamma(\boldsymbol{\beta},\eta,\gamma)\} + \gamma E\{(\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{S}^2(\boldsymbol{Y}-\boldsymbol{\mu})\} = \gamma E\{(\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{S}^2(\boldsymbol{Y}-\boldsymbol{\mu})\}$$

which equals 0 under H_0 and is positive under $H_1 : \gamma > 0$. So a large value of $U_\gamma(\boldsymbol{\beta},\eta,0)$ supports the alternative hypothesis. Furthermore, one can show that

$$\begin{cases} \boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta},\eta,\gamma) = \boldsymbol{X}^T(\boldsymbol{I}-\eta\boldsymbol{T}-\gamma\boldsymbol{S})(\boldsymbol{Y}-\boldsymbol{\mu}) = 0 \\ U_\eta(\boldsymbol{\beta},\eta,\gamma) = (\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{T}(\boldsymbol{I}-\eta\boldsymbol{T}-\gamma\boldsymbol{S})(\boldsymbol{Y}-\boldsymbol{\mu}) = 0 \end{cases}$$

are both unbiased by similar argument.

1.2 Asymptotic Representation of Q_G

We note that, for both GR and IBS similarity, $\boldsymbol{S}$ can be written as $\boldsymbol{Z}\boldsymbol{Z}^T + \boldsymbol{C}$, where $\boldsymbol{C} = -diag(\boldsymbol{Z}\boldsymbol{Z}^T)$ and is needed because in the definition of $\boldsymbol{S}$, subjects are not compared to themselves in terms of genetic similarity. For example, for GR similarity, $\boldsymbol{Z}(n \times q)$, is the centered genotype matrix, i.e., each column of the genotype matrix $\boldsymbol{G}$, $\boldsymbol{G}_{,h}$, is now centered by the genotype population mean $2p_h$, and for IBS similarity, $\boldsymbol{Z}$ is an $n \times 3q$ matrix again with each element defined in terms of genotype, described in the next subsection. Here we prove the following result.

Result 1.

$$Q_G = \frac{(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})^T\boldsymbol{S}(\boldsymbol{I}-\widehat{\eta}\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})}{m} = \frac{(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})^T\boldsymbol{Z}\boldsymbol{Z}^T(\boldsymbol{I}-\eta_0\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})}{m} + c + o_p(1),$$

where $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$; $\widehat{\eta}$ and $\widehat{\boldsymbol{\beta}}$ are the solution to estimating equations

$$\begin{cases} \boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta},\eta,0) = \boldsymbol{X}^T(\boldsymbol{I}-\eta\boldsymbol{T})(\boldsymbol{Y}-\boldsymbol{\mu}) = 0 \\ U_\eta(\boldsymbol{\beta},\eta,0) = (\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{T}(\boldsymbol{I}-\eta\boldsymbol{T})(\boldsymbol{Y}-\boldsymbol{\mu}) = 0. \end{cases}$$

Proof. We first note that

(B.1)

$$Q_G = \frac{(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})^T\boldsymbol{Z}\boldsymbol{Z}^T(\boldsymbol{I}-\widehat{\eta}\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})}{m} + \frac{1}{m}\sum_{i=1}^{m} c_i(\boldsymbol{Y}_i-\widehat{\boldsymbol{\mu}}_i)^T(\boldsymbol{I}_{n_i}-\widehat{\eta}\boldsymbol{T}_i)(\boldsymbol{Y}_i-\widehat{\boldsymbol{\mu}}_i),$$

where c_i is the (i,i)-th element of $\boldsymbol{C}$ which equals $-\sum_{h=1}^{q}(G_{i,h}-2p_h)^2$ for GR similarity and $-2q$ for IBS similarity; $\boldsymbol{T}_i$ is the (i,i)-th block of $\boldsymbol{T}$; $\boldsymbol{I}_{n_i}$ is an $n_i \times n_i$ identity matrix.

The first term in equation (B.1) is an inner product of $\frac{1}{\sqrt{m}}\boldsymbol{Z}^T(\boldsymbol{I}-\widehat{\eta}\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})$ and $\frac{1}{\sqrt{m}}\boldsymbol{Z}^T(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})$. We show that

$$\begin{aligned}
&\frac{1}{\sqrt{m}}\boldsymbol{Z}^T(\boldsymbol{I}-\widehat{\eta}\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})\\
=\ &\frac{1}{\sqrt{m}}\boldsymbol{Z}^T(\boldsymbol{I}-\eta_0\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})-\frac{1}{\sqrt{m}}\boldsymbol{Z}^T\boldsymbol{T}(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})(\widehat{\eta}-\eta_0)\\
=\ &\frac{1}{\sqrt{m}}\boldsymbol{Z}^T(\boldsymbol{I}-\eta_0\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})-\sqrt{m}(\widehat{\eta}-\eta_0)\{\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{Z}_i^T\boldsymbol{T}_i(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)+\boldsymbol{o}_p(1)\}\\
=\ &\frac{1}{\sqrt{m}}\boldsymbol{Z}^T(\boldsymbol{I}-\eta_0\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})+\boldsymbol{o}_p(1),\ \text{as}\ m\to\infty,
\end{aligned}$$

where the second equality follows by Taylor expansion at $\boldsymbol{\beta}$. Assuming that the number of repeated measurements of each subject is bounded, the estimator $\widehat{\eta}$ is $\sqrt{m}$-consistent for η under H_0 by the property of generalized estimating equations. Hence the last equality follows by the $\sqrt{m}$-consistency of $\widehat{\eta}$ and the weak law of large numbers. Therefore,

$$\frac{(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})^T\boldsymbol{Z}\boldsymbol{Z}^T(\boldsymbol{I}-\widehat{\eta}\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})}{m}=\frac{(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})^T\boldsymbol{Z}\boldsymbol{Z}^T(\boldsymbol{I}-\eta_0\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})}{m}+o_p(1).$$

Next we show that the second term in equation (B.1) asymptotically converges to a constant.

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}c_i(\boldsymbol{Y}_i-\widehat{\boldsymbol{\mu}}_i)^T(\boldsymbol{I}_{n_i}-\widehat{\eta}\boldsymbol{T}_i)(\boldsymbol{Y}_i-\widehat{\boldsymbol{\mu}}_i)\\
=\ &\frac{1}{m}\sum_{i=1}^{m}c_i(\boldsymbol{Y}_i-\widehat{\boldsymbol{\mu}}_i)^T(\boldsymbol{I}_{n_i}-\eta_0\boldsymbol{T}_i)(\boldsymbol{Y}_i-\widehat{\boldsymbol{\mu}}_i)-(\widehat{\eta}-\eta_0)\frac{1}{m}\sum_{i=1}^{m}c_i(\boldsymbol{Y}_i-\widehat{\boldsymbol{\mu}}_i)^T\boldsymbol{T}_i(\boldsymbol{Y}_i-\widehat{\boldsymbol{\mu}}_i)\\
=\ &\frac{1}{m}\sum_{i=1}^{m}c_i(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)^T(\boldsymbol{I}_{n_i}-\eta_0\boldsymbol{T}_i)(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)-(\widehat{\eta}-\eta_0)\frac{1}{m}\sum_{i=1}^{m}c_i(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)^T\boldsymbol{T}_i(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)+o_p(1)\\
=\ &c+o_p(1),
\end{aligned}$$

where the second equality is again by Taylor expansion and the last equality by the weak law of large numbers. Summarizing results, we have finished proving the result:

$$Q_G=\frac{(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})^T\boldsymbol{S}(\boldsymbol{I}-\widehat{\eta}\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})}{m}=\frac{(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})^T\boldsymbol{Z}\boldsymbol{Z}^T(\boldsymbol{I}-\eta_0\boldsymbol{T})(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})}{m}+c+o_p(1).$$

□

1.3 Robust Inference

We define $\widetilde{\boldsymbol{Z}}(\eta) = \{(\boldsymbol{I} - \eta\boldsymbol{T})\boldsymbol{Z}, \boldsymbol{Z}\}$, $\widetilde{\boldsymbol{X}}(\eta) = (\boldsymbol{I} - \eta\boldsymbol{T})\boldsymbol{X}$, and

$$\begin{aligned}\frac{1}{\sqrt{m}}\widetilde{\boldsymbol{R}}(\eta, \boldsymbol{\beta}) &= \frac{1}{\sqrt{m}}\left\{\begin{array}{c}\boldsymbol{R}_1(\eta, \boldsymbol{\beta})\\ \boldsymbol{R}_2(\eta, \boldsymbol{\beta})\end{array}\right\} = \frac{1}{\sqrt{m}}\left\{\begin{array}{c}\widetilde{\boldsymbol{Z}}(\eta)^T\\ \widetilde{\boldsymbol{X}}(\eta)^T\end{array}\right\}(\boldsymbol{Y} - \boldsymbol{\mu})\\ &= \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left\{\begin{array}{c}\widetilde{\boldsymbol{Z}}_i(\eta)^T\\ \widetilde{\boldsymbol{X}}_i(\eta)^T\end{array}\right\}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\widetilde{\boldsymbol{R}}_i(\eta, \boldsymbol{\beta}),\end{aligned}$$

where $\widetilde{\boldsymbol{Z}}_i(\eta)$, $\widetilde{\boldsymbol{X}}_i(\eta)$, $\boldsymbol{Y}_i$, $\boldsymbol{\mu}_i = \boldsymbol{X}_i^T\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{R}}_i(\eta, \boldsymbol{\beta})$ are the components corresponding to subject i respectively. We can rewrite $(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})^T\boldsymbol{Z}\boldsymbol{Z}^T(\boldsymbol{I} - \eta_0\boldsymbol{T})(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})/m$ as the quadratic form of $\frac{1}{\sqrt{m}}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}})$ by straightforward matrix algebra and have:

$$Q_G = \frac{1}{2m}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}})^T\begin{pmatrix}\boldsymbol{0}_{dq} & \boldsymbol{I}_{dq}\\ \boldsymbol{I}_{dq} & \boldsymbol{0}_{dq}\end{pmatrix}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}}) + c + o_p(1),$$

where $\boldsymbol{I}_{dq}$ is a $dq \times dq$ identity matrix and $\boldsymbol{0}_{dq}$ is a $dq \times dq$ matrix with all elements 0. In this subsection we prove the following and a directly followed results. We note that the proof does not rely on the normality assumption of outcomes $\boldsymbol{Y}$.

Result 2. Under the $H_0 : \gamma = 0$,

$$\frac{1}{\sqrt{m}}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}}) = \boldsymbol{A}\frac{1}{\sqrt{m}}\widetilde{\boldsymbol{R}}(\eta_0, \boldsymbol{\beta}_0) + \boldsymbol{o}_p(1) \Rightarrow N(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{A} = (\boldsymbol{I}_{2dq}, -E\{\frac{\partial \boldsymbol{R}_1(\eta_0,\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\}E^{-1}\{\frac{\partial \boldsymbol{R}_2(\eta_0,\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\})$, $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T$, and $\boldsymbol{D} = \text{var}\{\widetilde{\boldsymbol{R}}_i(\eta_0, \boldsymbol{\beta}_0)\}$. And $\boldsymbol{\Sigma}$ can be consistently estimated by the sandwich variance estimator $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{A}}^T$, where $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{D}}$ are the corresponding empirical counterpart defined below.

Proof. Note that $\widehat{\boldsymbol{\beta}}$ is the solution to $\boldsymbol{R}_2(\widehat{\eta}, \boldsymbol{\beta}) = 0$, i.e., $\frac{1}{\sqrt{m}}\boldsymbol{R}_2(\widehat{\eta}, \widehat{\boldsymbol{\beta}}) = 0$. We first show that $0 = \frac{1}{\sqrt{m}}\boldsymbol{R}_2(\widehat{\eta}, \widehat{\boldsymbol{\beta}}) = \frac{1}{\sqrt{m}}\boldsymbol{R}_2(\eta_0, \widehat{\boldsymbol{\beta}}) + \boldsymbol{o}_p(1)$. It follows because, by Taylor expansion,

$$\frac{1}{\sqrt{m}}\boldsymbol{R}_2(\widehat{\eta}, \widehat{\boldsymbol{\beta}}) \;=\; \frac{1}{\sqrt{m}}\boldsymbol{R}_2(\eta_0, \widehat{\boldsymbol{\beta}}) - \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{T}_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})\sqrt{m}(\widehat{\eta} - \eta_0),$$

and note $\sqrt{m}(\widehat{\eta} - \eta_0)$ is bounded in probability and $\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{T}_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})$ converges in probability to $E\{\boldsymbol{X}_i^T \boldsymbol{T}_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)\} = 0$, where $\boldsymbol{\beta}_0$ is the true parameter under H_0.

By Taylor expansion,

$$\begin{aligned} \frac{1}{\sqrt{m}}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}}) &= \frac{1}{\sqrt{m}}\boldsymbol{R}_1(\eta_0, \boldsymbol{\beta}_0) + \frac{1}{\sqrt{m}}\frac{\partial \boldsymbol{R}_1(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \boldsymbol{o}_p(1), \\ 0 = \frac{1}{\sqrt{m}}\boldsymbol{R}_2(\eta_0, \widehat{\boldsymbol{\beta}}) + \boldsymbol{o}_p(1) &= \frac{1}{\sqrt{m}}\boldsymbol{R}_2(\eta_0, \boldsymbol{\beta}_0) + \frac{1}{\sqrt{m}}\frac{\partial \boldsymbol{R}_2(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \boldsymbol{o}_p(1), \end{aligned}$$

Plugging the second equation into the first,

$$\begin{aligned} \frac{1}{\sqrt{m}}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}}) &= \frac{1}{\sqrt{m}}\boldsymbol{R}_1(\eta_0, \boldsymbol{\beta}_0) - \frac{\partial \boldsymbol{R}_1(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\{\frac{\partial \boldsymbol{R}_2(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\}^{-1}\frac{1}{\sqrt{m}}\boldsymbol{R}_2(\eta_0, \boldsymbol{\beta}_0) + \boldsymbol{o}_p(1) \\ &= (\boldsymbol{I}_{2dq}, -\frac{\partial \boldsymbol{R}_1(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\{\frac{\partial \boldsymbol{R}_2(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\}^{-1})\frac{1}{\sqrt{m}}\widetilde{\boldsymbol{R}}(\eta_0, \boldsymbol{\beta}_0) + \boldsymbol{o}_p(1). \end{aligned}$$

It is easy to see that

(B.2)
$$\frac{1}{m}\frac{\partial \boldsymbol{R}_1(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\{\frac{1}{m}\frac{\partial \boldsymbol{R}_2(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\}^{-1} = E\{\frac{\partial \boldsymbol{R}_1(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\}E^{-1}\{\frac{\partial \boldsymbol{R}_2(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\} + \boldsymbol{o}_p(1).$$

Thus

$$\begin{aligned} \frac{1}{\sqrt{m}}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}}) &= (\boldsymbol{I}_{2dq}, -E\{\frac{\partial \boldsymbol{R}_1(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\}E^{-1}\{\frac{\partial \boldsymbol{R}_2(\eta_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\})\frac{1}{\sqrt{m}}\widetilde{\boldsymbol{R}}(\eta_0, \boldsymbol{\beta}_0) + \boldsymbol{o}_p(1) \\ &= \boldsymbol{A}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\widetilde{\boldsymbol{R}}_i(\eta_0, \boldsymbol{\beta}_0) + \boldsymbol{o}_p(1), \end{aligned}$$

which by the central limit theory converges to a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T$. It is easy to check that, by the weak law of large numbers and the $\sqrt{m}$-consistency of $\widehat{\eta}$ and $\widehat{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}$ can be consistently estimated by the sandwich variance estimator $\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{A}}^T$, where

$$\widehat{\boldsymbol{D}} = \frac{1}{m}\sum_{i=1}^{m} \widetilde{\boldsymbol{R}}_i(\widehat{\eta}, \widehat{\boldsymbol{\beta}})\widetilde{\boldsymbol{R}}_i(\widehat{\eta}, \widehat{\boldsymbol{\beta}})^T,$$

$$\widehat{\boldsymbol{A}} = (\boldsymbol{I}_{2dq}, -\frac{\partial \boldsymbol{R}_1(\widehat{\eta}, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T}\{\frac{\partial \boldsymbol{R}_2(\widehat{\eta}, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T}\}^{-1}).$$

Specifically, $\frac{\boldsymbol{R}_1(\widehat{\eta}, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} = -\widetilde{\boldsymbol{Z}}(\widehat{\eta})^T\boldsymbol{X}$ and $\frac{\partial \boldsymbol{R}_2(\widehat{\eta}, \widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} = -\widetilde{\boldsymbol{X}}(\widehat{\eta})^T\boldsymbol{X}$. □

Result 3. Under regularity conditions, Q_G has an asymptotic distribution

$$\frac{1}{2}\sum_{i=1}^{2dq} \lambda_i \chi_i^2 + c, \tag{B.3}$$

where c is a constant which does not affect the inference; χ_i^2s are i.i.d. Chi-square distributions; λ_i are eigenvalues of a $2dq \times 2dq$ matrix

$$\begin{pmatrix} \mathbf{0}_{dq} & \boldsymbol{I}_{dq} \\ \boldsymbol{I}_{dq} & \mathbf{0}_{dq} \end{pmatrix} \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \\ \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \end{pmatrix};$$

$\boldsymbol{\Sigma}$ is defined above and can be consistently estimated by $\widehat{\boldsymbol{\Sigma}}$ as in Result 2; $\boldsymbol{\Sigma}_{11}$ is the first $dq \times dq$ block of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Sigma}_{22}$ can be defined similarly.

Proof. In the proof of Result 2, we have showed $\frac{1}{\sqrt{m}}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}}) \Rightarrow N(\mathbf{0}, \boldsymbol{\Sigma})$ under $H_0 : \gamma = 0$. Therefore,

$$Q_G = \frac{1}{2m}\boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}})^T \begin{pmatrix} \mathbf{0}_{dq} & \boldsymbol{I}_{dq} \\ \boldsymbol{I}_{dq} & \mathbf{0}_{dq} \end{pmatrix} \boldsymbol{R}_1(\eta_0, \widehat{\boldsymbol{\beta}}) + c + o_p(1)$$

is asymptotically distributed as

$$\frac{1}{2}\sum_{i=1}^{2dq} \lambda_i \chi_i^2 + c$$

by the property of quadratic form of normal random variables. In addition, $\boldsymbol{\Sigma}$ can be consistently estimated by $\widehat{\boldsymbol{\Sigma}}$ as we showed in Result 2.

□

1.4 Use of IBS metric

As discussed in the main paper, the Identity-by-state (IBS) similarity: $s_{i,j} = \rho(\boldsymbol{G}_i, \boldsymbol{G}_j) = \sum_{h=1}^{q}(2 - |G_{i,h} - G_{j,h}|)$, which has been commonly used, e.g., in SKAT, is an alternative choice to quantify genetic similarity in LGRF. However, the use of IBS kernel is limited by its computational inefficiency (Wu, et al., 2011), though they have recognized that IBS kernel usually has higher power than linear kernel in the presence of gene-gene interaction. We hereby propose a fast implementation of IBS metric in LGRF, and as a result both the robustness to misspecification of working correlation structure and computational efficiency can be achieved. Recall the genotype of a single genetic variant of subject i can be coded by $\{0,\,1,\,2\}$. We generate three pseudo-variables by

$$\begin{array}{r} 0: \\ 1: \\ 2: \end{array} \begin{pmatrix} \sqrt{2} & 0 & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 1 \\ 0 & \sqrt{2} & 0 \end{pmatrix} =: \boldsymbol{B}.$$

That is, the pseudo-variables are $(\sqrt{2}, 0, 0)$ if the genotype is 0; $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 1)$ if it is 1; $(0, \sqrt{2}, 0)$ if it is 2. The inner-product of two subjects' pseudo-variables exactly equal the IBS metric; that is,

$$\boldsymbol{B}\boldsymbol{B}^T = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

If we denote the pseudo-variables with respect to p genetic variants as $\boldsymbol{Z}_{IBS}$, an $n \times 3q$ matrix, the IBS metric between all pairs of subjects are $\boldsymbol{Z}_{IBS}\boldsymbol{Z}_{IBS}^T$. Therefore, genetic similarity in terms of the IBS metric can be represented as $\boldsymbol{S} = \boldsymbol{Z}_{IBS}\boldsymbol{Z}_{IBS}^T - \boldsymbol{C}$, where $\boldsymbol{C} = -diag(\boldsymbol{Z}_{IBS}\boldsymbol{Z}_{IBS}^T)$; again note the term $\boldsymbol{C}$ is due to that in the definition of $\boldsymbol{S}$, subjects are not compared to themselves in terms of genetic similarity. By using peudo-variables the computational efficiency increases dramatically, but is still slightly less compared with using the genetic relationship similarity because the number of variables increase from q to $3q$. This representation shows that the IBS metric corresponds to a linear model in SKAT with $3q$ pseudo variables, which actually does not model the interaction among genetic variants.

2. Additional Simulations

2.1 LGRF Run-time Simulation

To evaluate the computational performance of LGRF, we varied the number of total observations and recorded the running times for both fitting the null model and testing the target region C10orf107 (154 SNPs). The numbers of total observations were set to be 3000, 6000 and 10000, mimicking the total number of observations in CHN, HIS/AFA and CAU ethnic groups in MESA respectively. Supplementary Table B.1 shows the running times for testing a region with 154 SNPs on a 2.67GHz Linux PC with an Intel Xeon X5650 processor. The numbers of total observations (n) are 3000, 6000 and 10000, akin to those observed in the CHN, HIS/AFA and CAU ethnic groups respectively in MESA. For a longitudinal study containing 3000 observations, like the CHN ethnic group in MESA, LGRF-G requires 2.0 seconds to fit the null model and 0.6 seconds to calculate the p-value for the entire target region. Since the null model only need to be fit once, the computational cost for testing K regions is approximately $0.6 \times K$ seconds. We expect that the computational cost will increase if number of SNPs in the region increases, but this number is usually bounded by the length of the region. For example, the largest candidate region in our analysis has 1026 SNPs. The LGRF-J test requires longer time for calculating p-value because additional interaction terms are explicitly included. On the other hand, the running time increases as the number of observations increases. If the number of total observations is increased to 10000, such as the CAU ethnic in MESA, LGRF-G requires 9.7 seconds to fit the null model and 1.9 seconds to compute the p-value.

Table B.1: Running times corresponding to different number of total observations. The running times for both fitting the null model and testing the target region C10orf107 (190 SNPs) on a 2.67GHz Linux PC with an Intel Xeon X5650 processor are showed in this table. The numbers of total observations (n) are 3000, 6000 and 10000, approximating the numbers corresponding to CHN, HIS/AFA and CAU ethnic groups respectively, as observed in MESA. CAU: Caucasians; AFA: African Americans; HIS: Hispanics; CHN: Asians of Chinese descent.

Number of Total	Fitting the	Calculating the P-value	
Observations (n)	Null Model	LGRF-G	LGRF-J
3000	2.0 seconds	0.6 seconds	2.0 seconds
6000	4.9 seconds	1.1 seconds	3.4 seconds
10000	9.7 seconds	1.9 seconds	5.0 seconds

2.2 Simulations Investigating Meta/Mega-analysis Strategies with a Multi-Ethnic Cohort

We additionally simulated scenarios where four ethnic groups shared the same set of causal variants versus different set of causal variants in the same target region to compare meta and mega analysis, and show the advantage of gene-level meta-analysis over single-SNP meta-analysis. The gene-level meta-analysis evaluated the region for each race ethnicity by LGRF-G and combine the p-values by fisher's method. The single-SNP meta-analysis approach used the popular meta-analysis tool METAL proposed by Willer et al. (2010). Each SNP was tested using GEE-G within each ethnicity and the four Z-scores converted from the four ethnic groups' p-values were then combined to provide overall measures of significance. The minimum p-value was then adjusted for multiple testing by the Bonferroni correction. The mega-analysis was pooling the ethnic groups together and then applying LGRF-G and GEE-G using individual level data.

For each replicated dataset, four ethnic groups were randomly simulated from the CAU, AFA, HIS and CHN ethnic groups correspondingly, and gene region C10orf107 was chosen as the target region. The total number of subjects was 1000 and each subject had four repeated measurements. The sample sizes corresponding to the simulated ethnic groups were 400, 250, 220 and 130 respectively, proportional to those observed in MESA. A different SNP was randomly chosen to be causal within each ethnic group in the case of distinct effects, while four ethnic groups shared the same causal SNP in the common effect case. Specifically, the true model is of form:

$$Y_{i,l} = \alpha_0 t_{i,l} + \alpha_1 G_{E,i} + \epsilon_{i,l}, t_{i,l} = 1, \ldots, r,$$

where $G_{E,i}$ is the genotype of subject i for the randomly selected causal SNP of the ethnicity E that subject i belongs to; $\alpha_0 = 12/r$, $\alpha_1 = 0.4$; r is the number of measurements per subject. The missingness indicators and other simulation parameters were almost the same as the power simulation scenario I where we considered a single causal SNP that had a marginal effect and the within-subject correlation structure was CS.

Supplementary Table B.2 presents the comparisons. When the four ethnic groups have different causal variants, gene-level meta-analysis shows substantial higher power (0.832) than single-SNP meta-analysis (0.520). This is intuitive because single-SNP meta-analysis will dilute the signal of each causal variant as the strength is not uniform across each cohort at the SNP level. Moreover, a gene-level meta-analysis is preferred here than a mega-analysis using individual level data for the same reason that pooling the data together will dilute the signal. When the four ethnic groups have the same causal variant, gene-level meta-analysis achieves slightly lower power (0.724 vs. 0.782), because the signal was accumulated on the same variant while combining the four groups.

Table B.2: Power Studies for Meta/Mega-analysis when Causal Variants are Distinct/Common across Four Ethnic Groups. Each cell represents the empirical power from 500 replicates at level α=0.05. In each ethnic group, one randomly selected SNP is causal and has a marginal effect. LGRF-meta: use LGRF-G to test the region within ethnicity and combine the p-values by Fisher's method. LGRF-mega: jointly tests the four ethnic groups by pooling the individual level data. GEE-meta: use GEE-G to test each SNP within ethnicity and combine the p-values by METAL proposed by Willer, et al. (2010). GEE-mega: test each SNP using the individual level data of four ethnic groups jointly.

Causal Variants	Meta-analysis		Mega-analysis	
	LGRF-meta	GEE-meta	LGRF-mega	GEE-mega
Distinct across ethnic groups	0.832	0.520	0.614	0.558
Common across ethnic groups	0.724	0.782	0.754	0.820

2.3 Impact of Genetic Similarity Metrics

We evaluated the impact of the genetic similarity metrics in three main scenarios considered in table 3-5 of the main text and summarize the result in Supplementary Table B.3. The number of repeated measurements per subject is six, and there are 400 subjects in each replicate. The correlation structure among repeated measurements is compound symmetric. Detailed parameters are same as the three power simulation settings in the main text respectively. The IBS similarity has analogous performance as genetic relationship in the simulation studies considered in the paper.

Table B.3: Power Studies for evaluating the impact of the genetic similarity metric: Each cell represents the empirical power from 500 replicates at level α=0.05. The number of repeated measurements per subject is six, and there are 400 subjects in each replicate. The correlation structure among repeated measurements is compound symmetric. The parameter configurations are same as the three simulation settings (Tables 3-5) in the main text. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. Both LGRF and LGRF-J use the genetic similarity metric. LGRF-G-IBS: the LGRF test for the marginal effect of a gene using the identity-by state (IBS) similarity. LGRF-J-IBS: the LGRF test for the joint effect of gene and gene-time interaction. The genetic main effect is modeled using IBS similarity.

Simulation Scenario	Marginal Association Test		Joint Association Test	
	LGRF-G	LGRF-G-IBS	LGRF-J	LGRF-J-IBS
Single SNP Marginal Effect	0.426	0.438	0.417	0.364
Single SNP×Time effect	0.326	0.318	0.430	0.486
Multiple SNPs Combined Effect	0.330	0.324	0.354	0.344

2.4 Further Evaluation of the Power Difference

Given the analogous quadratic form of the LGRF score test and SKAT test, we expect they will have similar power if we have a longitudinal version of SKAT even if they are developed from two different perspective. To confirm this, we applied the LGRF test to the average of repeated measurements to three main scenarios considered in table 3-5 of the main text and present the result in Supplementary Table B.4. We also included the GenRF test for comparison. The number of repeated measurements per subject is six, and there are 400 subjects in each replicate. The correlation structure among repeated measurements is compound symmetric. Detailed parameters are same as the three simulation settings in the main text respectively. We observed that GenRF, SKAT and LGRF have comparable power in the scenarios considered here when they are all applied to the average of repeated measurements. This shows that the longitudinal design is the main reason of the power difference.

Table B.4: Power Studies including GenRF: Each cell represents the empirical power from 500 replicates at level α=0.05. The number of repeated measurements per subject is six, and there are 400 subjects in each replicate. The correlation structure among repeated measurements is compound symmetric. Detailed parameters are same as the three simulation settings (Table 3-5) in the main text respectively. LGRF-G: the LGRF test for the marginal effect of a gene using longitudinal dats. LGRF-Avg.: the LGRF test applied to the average of repeated measurements. GenRF-Avg.: the GenRF test applied to the average of repeated measurements. SKAT-Avg.: the SKAT test applied to the average of repeated measurements.

Causal Effect		Based on Average		
	LGRF-G	LGRF-Avg.	GenRF-Avg.	SKAT-Avg.
Single SNP Marginal Effect	0.426	0.290	0.287	0.290
Single SNP×Time effect	0.326	0.196	0.174	0.186
Multiple SNPs Combined Effect	0.330	0.214	0.196	0.208

2.5 Type - I Error Rate Evaluation at Lower Significance levels

We further evaluated LGRF-G (the LGRF test for the marginal effect of a gene) at a lower α level using 2.5×10^7 replicates. The smaller α level (2.5×10^{-6}) considered here reflects the scenario of a genome-wide gene-level analysis where we have approximately 20,000 genes in total. Other parameters are held same as the type I error simulations in the main text (with an α level of 0.001). We present the results in the Supplementary Table 5. As we expected, LGRF-G tends to be conservative in these scenarios due to the use of sandwich estimator as in regular GEE, which has been known to be slightly conservative.

Table B.5: Type-I error rate evaluation at small α level. Each cell represents the empirical type-I error rate of LGRF-G (the LGRF test for the marginal effect of a gene) based on 2.5×10^7 replicates. The total number of observations is 2,400 and repeated measurements per subject were generated in the same follow-up period according to different correlation structures. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. The working correlation assumed in LGRF-G is CS.

	Type-I Error Rate					
	Four Repeated Measurements (600 Subjects)					
$\alpha =$	0.05	10^{-3}	10^{-4}	10^{-5}	2.5×10^{-6}	10^{-6}
Ind.	0.0494	9.03×10^{-4}	7.79×10^{-5}	5.20×10^{-6}	9.20×10^{-7}	5.20×10^{-7}
CS	0.0493	8.98×10^{-4}	7.95×10^{-5}	5.20×10^{-6}	6.39×10^{-7}	4.33×10^{-7}
AR1	0.0494	8.98×10^{-4}	7.76×10^{-5}	5.12×10^{-6}	6.26×10^{-7}	4.63×10^{-7}

2.6 Power Evaluation at Lower Significance levels

We evaluated how power changes at $\alpha = 0.001$ and $\alpha = 2.5 \times 10^{-6}$ when the number of subjects increases from 1200 to 6000, and each subject has four repeated measurements. The α level considered here either approximates the scenario in our data analysis, in which we consider a replication study with 29 regions, or reflect the scenario of a genome-wide gene-level analysis where we have approximately 20,000 genes in total. The simulation scenario is similar to Table 5 in the main text, where 10 out of the 154 SNPs in the region were randomly set to be causal each time. Among them, six SNPs have only marginal effects, three have both marginal and interaction effects and the remaining one has only an interaction effect. The parameters are held same as Table 5 in the main text except the sample size, such that the total variation in the outcome explained by the SNPs (including gene-time interaction) is approximately 1.5% - 2.0%. We present the results in Supplementary Table 6 and 7. We observed that the relative power difference is similar to what we showed in Table 5 of the main text.

Table B.6: Power comparisons when the number of subjects ranges from 1,200 to 6,000 and four repeated measurements were recorded. Randomly selected multiple SNPs are causal and have both marginal and interaction effects. Each cell represents the empirical power from 500 replicates at level α=0.001. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction.

Power: Multiple SNPs Combined Effect					
			Ind.		
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
1200	0.18	0.17	0.10	0.08	0.06
1800	0.32	0.30	0.21	0.15	0.11
2400	0.40	0.42	0.26	0.25	0.20
3600	0.63	0.65	0.49	0.48	0.41
6000	0.82	0.82	0.71	0.71	0.65
			CS		
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
1200	0.24	0.24	0.17	0.11	0.08
1800	0.37	0.37	0.23	0.19	0.15
2400	0.50	0.53	0.38	0.32	0.29
3600	0.70	0.70	0.58	0.54	0.50
6000	0.83	0.84	0.77	0.76	0.73
			AR1		
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
1200	0.32	0.36	0.19	0.17	0.13
1800	0.52	0.54	0.35	0.32	0.28
2400	0.60	0.62	0.48	0.44	0.39
3600	0.80	0.84	0.65	0.67	0.65
6000	0.90	0.91	0.84	0.86	0.85
			RR		
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
1200	0.49	0.47	0.28	0.31	0.23
1800	0.62	0.62	0.44	0.47	0.40
2400	0.75	0.74	0.55	0.59	0.53
3600	0.86	0.85	0.75	0.81	0.77
6000	0.94	0.93	0.88	0.91	0.91

Table B.7: Power comparisons when the number of subjects ranges from 3,600 to 9,600 and four repeated measurements were recorded. Randomly selected multiple SNPs are causal and have both marginal and interaction effects. Each cell represents the empirical power from 500 replicates at level $\alpha = 2.5 \times 10^{-6}$. Ind.: the repeated measurements are independent. CS: the correlation is compound symmetric. AR1: the repeated measurements follow a first-order auto-regressive model. RR: observations follow a mixed model with a random intercept and a random slope. LGRF-G: the LGRF test for the marginal effect of a gene. LGRF-J: the LGRF test for the joint effect of gene and gene-time interaction. The working correlation assumed in LGRF is CS. SKAT-Avg.: cross-sectional SKAT using the average value of repeated measurements as the outcome. GEE-G: test the marginal association by GEE. GEE-J: jointly test the marginal association and gene-time interaction by GEE. These single-marker tests were implemented by testing every SNP in the region and adjusting the minimum p-value by the Bonferroni correction.

Power: Multiple SNPs Combined Effect					
			Ind.		
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
3600	0.25	0.23	0.17	0.19	0.15
4800	0.44	0.45	0.31	0.37	0.31
6000	0.51	0.49	0.35	0.41	0.37
7200	0.62	0.62	0.48	0.57	0.53
9600	0.78	0.78	0.65	0.74	0.70
			CS		
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
3600	0.31	0.32	0.21	0.26	0.22
4800	0.46	0.47	0.36	0.42	0.40
6000	0.58	0.57	0.47	0.53	0.51
7200	0.66	0.67	0.57	0.63	0.60
9600	0.80	0.79	0.74	0.78	0.76
			AR1		
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
3600	0.42	0.46	0.31	0.38	0.35
4800	0.60	0.64	0.44	0.55	0.52
6000	0.72	0.74	0.60	0.67	0.66
7200	0.77	0.78	0.68	0.76	0.76
9600	0.84	0.86	0.77	0.83	0.82
			RR		
	LGRF-G	LGRF-J	SKAT-Avg.	GEE-G	GEE-J
3600	0.58	0.57	0.42	0.54	0.50
4800	0.71	0.71	0.54	0.70	0.65
6000	0.81	0.80	0.69	0.80	0.78
7200	0.89	0.89	0.79	0.88	0.86
9600	0.89	0.90	0.84	0.91	0.89

3. Descriptive Statistics of MESA

Table B.8: Gender distribution of MESA subjects across site and race. Each cell represents the number of subject in the corresponding category. WFU: Wake Forest University, Winston Salem, NC; COL: Columbia University, New York, NY; JHU: Johns Hopkins University, Baltimore, MD; UMN: University of Minnesota, Twin Cities, MN; NWU: Northwestern University, Chicago, IL; UCLA: University of California - Lost Angeles, Los Angeles, CA.

	Gender		
Site	Female	Male	All
WFU	528	464	992
COL	536	434	970
JHU	556	488	1044
UMN	532	518	1050
NWU	551	508	1059
UCLA	666	648	1314
All	3369	3060	6429

	Gender		
Race	Female	Male	All
White/Caucasian	1321	1206	2527
Chinese American	394	381	775
Black/African-American	906	771	1677
Hispanic	748	702	1450
All	3369	3060	6429

Table B.9: Longitudinal summary of blood pressure phenotypes and covariates we adjusted for in MESA across four exams. Sd: standard deviation. N: number of subject. sBP: systolic blood pressure. dBP: diastolic blood pressure. BMI: body mass index.

	Exam 1 (24 months)			Exam 2 (18 months)			Exam 3 (18 months)			Exam 4 (24 months)		
	Mean	Sd	N	Mean	Sd	N	Mean	Sd	N	Mean	Sd	N
sBP (mm Hg)	126.51	21.55	6427	124.33	20.79	5898	123.16	20.58	5619	123.59	20.56	5399
dBP (mm Hg)	71.82	10.27	6427	70.37	10.09	5898	69.69	9.94	5619	69.61	10.05	5399
BMI (kg/m^2)	28.3	5.47	6429	28.33	5.48	5889	28.28	5.51	5621	28.38	5.58	5402
Age (years)	62.22	10.24	6429	63.69	10.1	5900	64.99	9.99	5628	66.51	9.94	5505

Table B.10: Sensitivity analysis of the top 5 principal components (PCs) in MESA. Each cell represents the p-value. The analysis was done by fitting a multivariate linear regression using exam 1 data in MESA.

Race	PC1	PC2	PC3	PC4	PC5
White/Caucasian	0.0002	0.3502	0.2895	0.8484	0.0315
Chinese American	< 0.0001	0.0439	0.7882	0.9913	0.1982
Black/African-American	0.0029	0.3976	0.1406	0.3066	0.3808
Hispanic	0.9428	0.0306	0.4939	0.1760	0.6313

Table B.11: Chromosomal Region Information for the 29 regions considered in the MESA analysis.

Region Name	Chromosome	Start	End	Index SNP	Nearest Gene	Coded Allele Frequency
MOV10	1	113012286	113049891	rs2932538	MOV10	0.75
rs13082711	3	27462913	27562913	rs13082711	SLC4A7	0.78
MECOM	3	170278981	170869100	rs419076	MECOM	0.47
SLC39A8	4	103386221	103576438	rs13107325	SLC39A8	0.05
GUCY1A3	4	156802313	156877951	rs13139571	GUCY1A3,GUCY1B3	0.76
rs1173771	5	32800785	32900785	rs1173771	NPR3,C5orf23	0.6
rs11953630	5	157727980	157827980	rs11953630	EBF1	0.37
HFE	6	26190488	26211550	rs1799945	HFE	0.14
rs805303	6	31674345	31774345	rs805303	BAT2,BAT5	0.61
rs4373814	10	18409978	18509978	rs4373814	CACNB2	0.55
PLCE1	10	95738736	96083139	rs932764	PLCE1	0.44
rs7129220	11	10257114	10357114	rs7129220	ADM	0.89
ARHGAP42	11	100058594	100371866	rs633185	FLJ32810,TMEM133	0.28
FES	15	89222929	89245010	rs2521501	FURIN,FES	0.31
GOSR2	17	42350482	42465002	rs17608766	GOSR2	0.86
rs1327235	20	10867030	10967030	rs1327235	JAG1	0.46
rs6015450	20	57134512	57234512	rs6015450	GNAS,EDN3	0.12
MTHFR	1	11763367	11794564	rs17367504	MTHFR,NPPB	0.15
ULK4	3	41258094	41983926	rs3774372	ULK4	0.83
rs1458038	4	81333747	81433747	rs1458038	FGF5	0.29
CACNB2	10	18464612	18875804	rs1813353	CACNB2	0.68
C10orf107	10	63087725	63201530	rs4590817	C10orf107	0.84
NT5C2	10	104830930	104948046	rs11191548	CYP17A1,NT5C2	0.91
PLEKHA7	11	16751418	16997566	rs381815	PLEKHA7	0.26
ATP2B1	12	88500959	88632208	rs17249754	ATP2B1	0.84
SH2B3	12	110323135	110378810	rs3184504	SH2B3	0.47
rs10850411	12	113822179	113922179	rs10850411	TBX5,TBX3	0.7
rs1378942	15	72814420	72914420	rs1378942	CYP1A1,ULK3	0.35
ZNF652	17	44716567	44799834	rs12940887	ZNF652	0.38

4. Detailed Data Analysis of MESA

This section reports the full data analysis results of analyzing 29 candidate regions using the data from MESA. In addition to LGRF tests and SKAT, we also carried out an individual SNP based analysis (MinP) by testing every SNP in the region using GEE and adjusting the minimum p-value for multiple testing correction by multiplying with the effective number of independent tests explaining 99.95% variation (Gao, Starmer, and Martin, 2008). This proportion was determined by simulation such that the type-I error rate is neither inflated nor conservative (data not shown). We note that the preservation of nominal type-I error levels cannot be ensured because the real data scenarios can be very different from the simulated data due to the various effect sizes, proportions of causal variants, LD structures and so on. Supplementary Table B.12 - B.19 show the results of multi-ethnic groups analysis and table C.7 - C.8 show the results of meta-analysis.

Table B.12: CAU ethnic group sBP (2526 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements.

	name	chr	start	end	# of SNPs	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	42	0.13535	0.07635	0.09202	0.23265
2	MTHFR	1	11763367	11794564	72	0.01945	0.05276	0.05533	0.12314
3	rs13082711	3	27462913	27562913	111	0.01381	0.00466	0.00520	0.00778
4	MECOM	3	170278981	170869100	1026	1.00000	0.06022	0.10837	0.17882
5	ULK4	3	41258094	41983926	1000	0.22846	0.71848	0.70611	0.71814
6	SLC39A8	4	103386221	103576438	249	1.00000	0.86085	0.80668	0.81795
7	GUCY1A3	4	156802313	156877951	134	1.00000	0.53377	0.44202	0.33637
8	rs1458038	4	81333747	81433747	145	0.96055	0.22415	0.10866	0.05626
9	rs1173771	5	32800785	32900785	212	1.00000	0.33519	0.43248	0.54914
10	rs11953630	5	157727980	157827980	181	1.00000	0.89904	0.71726	0.86028
11	HFE	6	26190488	26211550	36	0.54141	0.72512	0.71851	0.68847
12	rs805303	6	31674345	31774345	146	0.00190	0.13020	0.08986	0.07329
13	rs4373814	10	18409978	18509978	284	0.59315	0.17728	0.20288	0.59591
14	PLCE1	10	95738736	96083139	401	1.00000	0.58603	0.61264	0.82326
15	CACNB2	10	18464612	18875804	902	0.38080	0.07818	0.10274	0.18348
16	C10orf107	10	63087725	63201530	190	1.00000	0.88840	0.92763	0.97798
17	NT5C2	10	104830930	104948046	113	0.82815	0.07243	0.08205	0.12620
18	rs7129220	11	10257114	10357114	178	1.00000	0.36771	0.45173	0.66651
19	ARHGAP42	11	100058594	100371866	715	1.00000	0.78860	0.71982	0.82962
20	PLEKHA7	11	16751418	16997566	464	0.16019	0.16659	0.17554	0.19694
21	ATP2B1	12	88500959	88632208	169	1.00000	0.22803	0.29551	0.56748
22	SH2B3	12	110323135	110378810	45	0.48835	0.48136	0.34815	0.53985
23	rs10850411	12	113822179	113922179	260	1.00000	0.43682	0.44055	0.51080
24	FES	15	89222929	89245010	18	1.00000	0.18458	0.28370	0.54740
25	rs1378942	15	72814420	72914420	84	0.00797	0.00186	0.00185	0.00233
26	GOSR2	17	42350482	42465002	138	0.37358	0.34244	0.37069	0.35589
27	ZNF652	17	44716567	44799834	79	1.00000	0.49696	0.57930	0.61687
28	rs1327235	20	10867030	10967030	313	0.21181	0.30059	0.28128	0.33724
29	rs6015450	20	57134512	57234512	180	0.27332	0.61897	0.57733	0.66472

Table B.13: CAU ethnic group dBP (2526 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements.

	name	chr	start	end	# of SNPs	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	42	0.46110	0.30341	0.27777	0.46583
2	MTHFR	1	11763367	11794564	72	0.00661	0.00130	0.00103	0.00355
3	rs13082711	3	27462913	27562913	111	0.31048	0.17037	0.17740	0.11849
4	MECOM	3	170278981	170869100	1026	0.04833	0.00513	0.02790	0.02207
5	ULK4	3	41258094	41983926	1000	0.04744	0.96234	0.91018	0.88880
6	SLC39A8	4	103386221	103576438	249	0.34061	0.55479	0.53973	0.28137
7	GUCY1A3	4	156802313	156877951	134	1.00000	0.47233	0.56610	0.87517
8	rs1458038	4	81333747	81433747	145	1.00000	0.29768	0.18177	0.25125
9	rs1173771	5	32800785	32900785	212	1.00000	0.66696	0.85676	0.89440
10	rs11953630	5	157727980	157827980	181	0.63602	0.23363	0.26583	0.11110
11	HFE	6	26190488	26211550	36	0.54645	0.34051	0.44308	0.45808
12	rs805303	6	31674345	31774345	146	0.30204	0.06300	0.04414	0.05236
13	rs4373814	10	18409978	18509978	284	0.16493	0.29760	0.32476	0.62033
14	PLCE1	10	95738736	96083139	401	1.00000	0.80312	0.82058	0.79770
15	CACNB2	10	18464612	18875804	902	0.08906	0.16648	0.14358	0.15338
16	C10orf107	10	63087725	63201530	190	0.02939	0.02024	0.02833	0.04118
17	NT5C2	10	104830930	104948046	113	1.00000	0.33286	0.25671	0.28467
18	rs7129220	11	10257114	10357114	178	0.66707	0.07883	0.08443	0.30574
19	ARHGAP42	11	100058594	100371866	715	1.00000	0.71191	0.54805	0.68127
20	PLEKHA7	11	16751418	16997566	464	0.61074	0.07355	0.08251	0.13227
21	ATP2B1	12	88500959	88632208	169	0.43572	0.75429	0.75579	0.61656
22	SH2B3	12	110323135	110378810	45	0.04899	0.16641	0.11498	0.21386
23	rs10850411	12	113822179	113922179	260	1.00000	0.35572	0.32875	0.36983
24	FES	15	89222929	89245010	18	1.00000	0.70865	0.74509	0.82268
25	rs1378942	15	72814420	72914420	84	0.25829	0.04784	0.02814	0.03862
26	GOSR2	17	42350482	42465002	138	1.00000	0.82360	0.99022	0.99637
27	ZNF652	17	44716567	44799834	79	0.96038	0.65212	0.85261	0.84157
28	rs1327235	20	10867030	10967030	313	1.00000	0.40429	0.40808	0.57369
29	rs6015450	20	57134512	57234512	180	0.39288	0.10504	0.09364	0.16543

Table B.14: AFA ethnic group sBP (1611 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements.

	name	chr	start	end	# of SNPs	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	25	0.41614	0.29182	0.24842	0.31362
2	MTHFR	1	11763367	11794564	60	1.00000	0.74027	0.76483	0.76851
3	rs13082711	3	27462913	27562913	82	1.00000	0.68060	0.67502	0.63146
4	MECOM	3	170278981	170869100	841	1.00000	0.92017	0.91854	0.95864
5	ULK4	3	41258094	41983926	740	1.00000	0.93804	0.92316	0.82468
6	SLC39A8	4	103386221	103576438	218	1.00000	0.56093	0.39974	0.67028
7	GUCY1A3	4	156802313	156877951	111	0.61661	0.05508	0.05925	0.07713
8	rs1458038	4	81333747	81433747	122	1.00000	0.42951	0.50710	0.42752
9	rs1173771	5	32800785	32900785	167	1.00000	0.97818	0.95217	0.96964
10	rs11953630	5	157727980	157827980	153	0.81745	0.25415	0.28025	0.41883
11	HFE	6	26190488	26211550	32	0.12438	0.03934	0.04456	0.08545
12	rs805303	6	31674345	31774345	132	1.00000	0.20036	0.19171	0.15740
13	rs4373814	10	18409978	18509978	235	1.00000	0.90417	0.88062	0.80690
14	PLCE1	10	95738736	96083139	313	0.00837	0.16311	0.31370	0.13983
15	CACNB2	10	18464612	18875804	741	1.00000	0.92649	0.85592	0.92046
16	C10orf107	10	63087725	63201530	157	1.00000	0.54669	0.55742	0.46793
17	NT5C2	10	104830930	104948046	89	0.12760	0.00497	0.00747	0.01635
18	rs7129220	11	10257114	10357114	147	1.00000	0.92541	0.90917	0.89739
19	ARHGAP42	11	100058594	100371866	580	0.86629	0.55717	0.40753	0.57759
20	PLEKHA7	11	16751418	16997566	386	0.20144	0.12597	0.12380	0.06895
21	ATP2B1	12	88500959	88632208	127	0.99723	0.33890	0.32429	0.34161
22	SH2B3	12	110323135	110378810	39	0.57348	0.12702	0.09567	0.19792
23	rs10850411	12	113822179	113922179	214	1.00000	0.94877	0.89189	0.84952
24	FES	15	89222929	89245010	14	0.64001	0.60704	0.59015	0.51314
25	rs1378942	15	72814420	72914420	70	1.00000	0.19290	0.18940	0.20472
26	GOSR2	17	42350482	42465002	121	1.00000	0.82218	0.80143	0.89642
27	ZNF652	17	44716567	44799834	60	1.00000	0.97062	0.96289	0.90895
28	rs1327235	20	10867030	10967030	187	1.00000	0.70146	0.70000	0.64716
29	rs6015450	20	57134512	57234512	101	0.31179	0.20503	0.23915	0.29266

Table B.15: AFA ethnic group dBP (1611 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements.

	name	chr	start	end	# of SNPs	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	25	0.64720	0.61927	0.53021	0.64604
2	MTHFR	1	11763367	11794564	60	0.88212	0.87692	0.96197	0.97343
3	rs13082711	3	27462913	27562913	82	0.33043	0.02325	0.02634	0.02223
4	MECOM	3	170278981	170869100	841	0.97042	0.59415	0.47938	0.43362
5	ULK4	3	41258094	41983926	740	0.65391	0.15723	0.21763	0.50748
6	SLC39A8	4	103386221	103576438	218	1.00000	0.62477	0.52559	0.77894
7	GUCY1A3	4	156802313	156877951	111	1.00000	0.08871	0.08911	0.16266
8	rs1458038	4	81333747	81433747	122	1.00000	0.31844	0.24816	0.17665
9	rs1173771	5	32800785	32900785	167	1.00000	0.76603	0.76637	0.81074
10	rs11953630	5	157727980	157827980	153	0.28363	0.28390	0.42828	0.49530
11	HFE	6	26190488	26211550	32	0.39060	0.08901	0.12339	0.10605
12	rs805303	6	31674345	31774345	132	0.05749	0.06843	0.08120	0.03396
13	rs4373814	10	18409978	18509978	235	1.00000	0.97314	0.80425	0.61203
14	PLCE1	10	95738736	96083139	313	1.00000	0.43693	0.44077	0.38918
15	CACNB2	10	18464612	18875804	741	0.54101	0.88589	0.82264	0.91039
16	C10orf107	10	63087725	63201530	157	0.09040	0.01524	0.01296	0.01055
17	NT5C2	10	104830930	104948046	89	1.00000	0.39532	0.41655	0.65896
18	rs7129220	11	10257114	10357114	147	0.26467	0.43629	0.36792	0.29269
19	ARHGAP42	11	100058594	100371866	580	0.58105	0.14785	0.26547	0.42156
20	PLEKHA7	11	16751418	16997566	386	1.00000	0.51775	0.59301	0.29514
21	ATP2B1	12	88500959	88632208	127	1.00000	0.09377	0.11849	0.14564
22	SH2B3	12	110323135	110378810	39	0.14582	0.58411	0.65584	0.25757
23	rs10850411	12	113822179	113922179	214	0.26957	0.86843	0.87413	0.90939
24	FES	15	89222929	89245010	14	1.00000	0.66729	0.64717	0.43090
25	rs1378942	15	72814420	72914420	70	1.00000	0.47473	0.49868	0.56505
26	GOSR2	17	42350482	42465002	121	0.39961	0.12266	0.17179	0.38108
27	ZNF652	17	44716567	44799834	60	1.00000	0.20299	0.21907	0.16900
28	rs1327235	20	10867030	10967030	187	1.00000	0.51377	0.72879	0.71563
29	rs6015450	20	57134512	57234512	101	0.32563	0.16282	0.17482	0.17651

Table B.16: HIS ethnic group sBP (1449 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements.

	name	chr	start	end	# of SNPs	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	25	0.21720	0.19764	0.14669	0.22319
2	MTHFR	1	11763367	11794564	60	1.00000	0.30060	0.26727	0.05594
3	rs13082711	3	27462913	27562913	82	0.30355	0.03073	0.02657	0.04527
4	MECOM	3	170278981	170869100	841	0.19305	0.13142	0.12776	0.04830
5	ULK4	3	41258094	41983926	740	1.00000	0.82240	0.85875	0.63704
6	SLC39A8	4	103386221	103576438	218	1.00000	0.96378	0.97076	0.94457
7	GUCY1A3	4	156802313	156877951	111	0.05460	0.16260	0.19888	0.27068
8	rs1458038	4	81333747	81433747	122	0.18943	0.37949	0.31693	0.41744
9	rs1173771	5	32800785	32900785	167	0.06474	0.11461	0.06462	0.21419
10	rs11953630	5	157727980	157827980	153	1.00000	0.27397	0.31941	0.38667
11	HFE	6	26190488	26211550	32	0.94388	0.42017	0.28997	0.26090
12	rs805303	6	31674345	31774345	132	0.30084	0.32730	0.28558	0.45846
13	rs4373814	10	18409978	18509978	235	0.07423	0.05311	0.03130	0.03711
14	PLCE1	10	95738736	96083139	313	0.50239	0.13823	0.11782	0.06340
15	CACNB2	10	18464612	18875804	742	0.27413	0.12414	0.07441	0.09030
16	C10orf107	10	63087725	63201530	157	0.00208	0.00795	0.00330	0.00233
17	NT5C2	10	104830930	104948046	89	1.00000	0.49682	0.39424	0.59208
18	rs7129220	11	10257114	10357114	147	0.02707	0.21695	0.21966	0.25168
19	ARHGAP42	11	100058594	100371866	580	0.69250	0.45838	0.30851	0.13242
20	PLEKHA7	11	16751418	16997566	387	0.64844	0.25484	0.22700	0.08798
21	ATP2B1	12	88500959	88632208	127	1.00000	0.47628	0.68267	0.65226
22	SH2B3	12	110323135	110378810	39	1.00000	0.95814	0.95237	0.98013
23	rs10850411	12	113822179	113922179	214	1.00000	0.70291	0.56957	0.51005
24	FES	15	89222929	89245010	14	0.47022	0.28032	0.32165	0.34811
25	rs1378942	15	72814420	72914420	70	0.95797	0.40943	0.52693	0.34465
26	GOSR2	17	42350482	42465002	121	0.25193	0.50482	0.45614	0.48149
27	ZNF652	17	44716567	44799834	60	1.00000	0.39857	0.32511	0.48670
28	rs1327235	20	10867030	10967030	187	1.00000	0.55602	0.65414	0.61712
29	rs6015450	20	57134512	57234512	101	1.00000	0.62157	0.55568	0.18535

Table B.17: HIS ethnic group dBP (1449 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements.

	name	chr	start	end	# of SNPs	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	25	1.00000	0.74680	0.87073	0.63924
2	MTHFR	1	11763367	11794564	60	0.34353	0.08581	0.05394	0.02890
3	rs13082711	3	27462913	27562913	82	0.00459	0.00577	0.00861	0.03495
4	MECOM	3	170278981	170869100	841	0.33805	0.22671	0.25818	0.20534
5	ULK4	3	41258094	41983926	740	1.00000	0.90854	0.95999	0.90103
6	SLC39A8	4	103386221	103576438	218	1.00000	0.53937	0.36558	0.07364
7	GUCY1A3	4	156802313	156877951	111	0.11861	0.30924	0.36171	0.40988
8	rs1458038	4	81333747	81433747	122	0.47361	0.25443	0.24657	0.25349
9	rs1173771	5	32800785	32900785	167	0.02616	0.54847	0.38889	0.71940
10	rs11953630	5	157727980	157827980	153	1.00000	0.36324	0.37470	0.46645
11	HFE	6	26190488	26211550	32	1.00000	0.86865	0.92151	0.93876
12	rs805303	6	31674345	31774345	132	0.16574	0.37206	0.39796	0.52797
13	rs4373814	10	18409978	18509978	235	1.00000	0.47485	0.32155	0.39683
14	PLCE1	10	95738736	96083139	313	1.00000	0.54681	0.53568	0.41218
15	CACNB2	10	18464612	18875804	742	0.88106	0.32142	0.24189	0.51498
16	C10orf107	10	63087725	63201530	157	0.05212	0.02343	0.01039	0.00814
17	NT5C2	10	104830930	104948046	89	1.00000	0.89105	0.76728	0.60572
18	rs7129220	11	10257114	10357114	147	0.12894	0.04984	0.06132	0.07062
19	ARHGAP42	11	100058594	100371866	580	0.74441	0.13852	0.09680	0.15075
20	PLEKHA7	11	16751418	16997566	387	0.10437	0.60733	0.49726	0.23367
21	ATP2B1	12	88500959	88632208	127	0.24094	0.40082	0.45522	0.42751
22	SH2B3	12	110323135	110378810	39	1.00000	0.96279	0.96176	0.98044
23	rs10850411	12	113822179	113922179	214	1.00000	0.73359	0.55539	0.52110
24	FES	15	89222929	89245010	14	1.00000	0.80443	0.75815	0.84079
25	rs1378942	15	72814420	72914420	70	0.17513	0.40532	0.29054	0.17715
26	GOSR2	17	42350482	42465002	121	0.11577	0.52576	0.53297	0.42869
27	ZNF652	17	44716567	44799834	60	1.00000	0.45140	0.37263	0.64254
28	rs1327235	20	10867030	10967030	187	1.00000	0.77211	0.87579	0.94684
29	rs6015450	20	57134512	57234512	101	1.00000	0.86702	0.91764	0.80125

Table B.18: CHN ethnic group sBP (775 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements.

	name	chr	start	end	# of SNPs	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	26	0.15059	0.36309	0.28488	0.24989
2	MTHFR	1	11763367	11794564	59	1.00000	0.60830	0.55853	0.71237
3	rs13082711	3	27462913	27562913	79	0.10178	0.03019	0.01908	0.04958
4	MECOM	3	170278981	170869100	822	0.64821	0.39291	0.54738	0.75053
5	ULK4	3	41258094	41983926	722	1.00000	0.44015	0.36834	0.49278
6	SLC39A8	4	103386221	103576438	213	1.00000	0.56075	0.50357	0.71217
7	GUCY1A3	4	156802313	156877951	107	0.22165	0.48432	0.46386	0.40801
8	rs1458038	4	81333747	81433747	120	0.36911	0.18223	0.24495	0.27591
9	rs1173771	5	32800785	32900785	159	1.00000	0.36743	0.18806	0.18593
10	rs11953630	5	157727980	157827980	150	1.00000	0.97636	0.94505	0.85109
11	HFE	6	26190488	26211550	31	0.59685	0.20983	0.26259	0.33061
12	rs805303	6	31674345	31774345	130	1.00000	0.88966	0.82054	0.73643
13	rs4373814	10	18409978	18509978	229	1.00000	0.31388	0.33109	0.39644
14	PLCE1	10	95738736	96083139	305	1.00000	0.72016	0.63652	0.82586
15	CACNB2	10	18464612	18875804	714	0.26030	0.02177	0.02603	0.00944
16	C10orf107	10	63087725	63201530	154	1.00000	0.46945	0.46612	0.38714
17	NT5C2	10	104830930	104948046	88	0.01400	0.81579	0.85689	0.93057
18	rs7129220	11	10257114	10357114	142	1.00000	0.50385	0.49822	0.32532
19	ARHGAP42	11	100058594	100371866	554	0.02062	0.16038	0.13582	0.17357
20	PLEKHA7	11	16751418	16997566	378	0.41303	0.24477	0.14763	0.27488
21	ATP2B1	12	88500959	88632208	122	0.63404	0.39300	0.36684	0.18625
22	SH2B3	12	110323135	110378810	36	0.21587	0.37482	0.41033	0.42779
23	rs10850411	12	113822179	113922179	211	0.89265	0.97007	0.92754	0.90111
24	FES	15	89222929	89245010	14	1.00000	0.22070	0.22339	0.32847
25	rs1378942	15	72814420	72914420	70	1.00000	0.89694	0.87975	0.93642
26	GOSR2	17	42350482	42465002	116	0.99234	0.17330	0.15325	0.23567
27	ZNF652	17	44716567	44799834	59	0.39036	0.24913	0.47506	0.58936
28	rs1327235	20	10867030	10967030	184	1.00000	0.20434	0.18397	0.24773
29	rs6015450	20	57134512	57234512	96	0.18157	0.07655	0.08302	0.08470

Table B.19: CHN ethnic group dBP (775 subjects) Analysis. Four ethnic groups were analyzed separately using LGRF, SKAT and MinP. The analysis was done under the adjustment of age, gender, BMI and top two principle components to correct for potential within-ethnicity stratification. MinP: testing every SNP with GEE in the region and adjusting the minimum p-value by the effective number of variants explaining 99.95% of the genotype variation in the region. SKAT was applied to the average value of repeated measurements.

	name	chr	start	end	# of SNPs	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	26	0.17209	0.31239	0.26490	0.37278
2	MTHFR	1	11763367	11794564	59	1.00000	0.40381	0.36832	0.35052
3	rs13082711	3	27462913	27562913	79	0.03003	0.03078	0.02917	0.07132
4	MECOM	3	170278981	170869100	822	1.00000	0.48813	0.40892	0.44618
5	ULK4	3	41258094	41983926	722	0.06193	0.24029	0.18058	0.37429
6	SLC39A8	4	103386221	103576438	213	1.00000	0.78774	0.60471	0.68886
7	GUCY1A3	4	156802313	156877951	107	0.97396	0.29995	0.31670	0.38890
8	rs1458038	4	81333747	81433747	120	1.00000	0.41763	0.58483	0.58023
9	rs1173771	5	32800785	32900785	159	1.00000	0.40452	0.33038	0.29395
10	rs11953630	5	157727980	157827980	150	1.00000	0.99585	0.88124	0.85264
11	HFE	6	26190488	26211550	31	0.44131	0.13491	0.18534	0.32330
12	rs805303	6	31674345	31774345	130	0.53269	0.22116	0.21903	0.49052
13	rs4373814	10	18409978	18509978	229	0.83391	0.33075	0.30770	0.37043
14	PLCE1	10	95738736	96083139	305	0.90713	0.97003	0.93120	0.89614
15	CACNB2	10	18464612	18875804	714	0.12518	0.01679	0.01644	0.00552
16	C10orf107	10	63087725	63201530	154	1.00000	0.47566	0.53613	0.49981
17	NT5C2	10	104830930	104948046	88	1.00000	0.76354	0.70835	0.74797
18	rs7129220	11	10257114	10357114	142	1.00000	0.60813	0.64742	0.53630
19	ARHGAP42	11	100058594	100371866	554	0.10876	0.58452	0.58966	0.77740
20	PLEKHA7	11	16751418	16997566	378	1.00000	0.14965	0.08344	0.17324
21	ATP2B1	12	88500959	88632208	122	0.24554	0.21366	0.20688	0.06752
22	SH2B3	12	110323135	110378810	36	0.92189	0.73576	0.65778	0.61232
23	rs10850411	12	113822179	113922179	211	1.00000	0.80558	0.98446	0.98229
24	FES	15	89222929	89245010	14	1.00000	0.09102	0.09950	0.16090
25	rs1378942	15	72814420	72914420	70	1.00000	0.59380	0.70537	0.73863
26	GOSR2	17	42350482	42465002	116	0.82658	0.18860	0.18656	0.18634
27	ZNF652	17	44716567	44799834	59	0.23437	0.84935	0.94165	0.94194
28	rs1327235	20	10867030	10967030	184	1.00000	0.85964	0.78668	0.82811
29	rs6015450	20	57134512	57234512	96	0.38382	0.29070	0.35086	0.32896

Table B.20: Meta-Analysis of sBP (6361 subjects) in MESA. Four ethnic groups were combined using Fisher's method.

	name	chr	start	end	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	0.12662	0.11615	0.08422	0.20122
2	MTHFR	1	11763367	11794564	0.44532	0.27329	0.25608	0.19276
3	rs13082711	3	27462913	27562913	0.04979	0.00129	0.00087	0.00359
4	MECOM	3	170278981	170869100	0.84271	0.16447	0.26964	0.25385
5	ULK4	3	41258094	41983926	0.93729	0.94506	0.96538	0.90941
6	SLC39A8	4	103386221	103576438	1.00000	0.95245	0.88354	0.98117
7	GUCY1A3	4	156802313	156877951	0.27964	0.14522	0.14894	0.16461
8	rs1458038	4	81333747	81433747	0.71395	0.26335	0.20692	0.16141
9	rs1173771	5	32800785	32900785	0.70582	0.38030	0.22572	0.46261
10	rs11953630	5	157727980	157827980	0.99994	0.69307	0.69145	0.83252
11	HFE	6	26190488	26211550	0.58656	0.15250	0.14972	0.22746
12	rs805303	6	31674345	31774345	0.06037	0.28225	0.20031	0.19631
13	rs4373814	10	18409978	18509978	0.61972	0.15804	0.12702	0.27193
14	PLCE1	10	95738736	96083139	0.20492	0.31685	0.38811	0.24973
15	CACNB2	10	18464612	18875804	0.51404	0.02932	0.02661	0.02366
16	C10orf107	10	63087725	63201530	0.13607	0.12543	0.07493	0.04878
17	NT5C2	10	104830930	104948046	0.11075	0.02387	0.03046	0.09402
18	rs7129220	11	10257114	10357114	0.51322	0.58221	0.62433	0.64354
19	ARHGAP42	11	100058594	100371866	0.36079	0.55124	0.35965	0.34083
20	PLEKHA7	11	16751418	16997566	0.30172	0.10267	0.07077	0.04177
21	ATP2B1	12	88500959	88632208	0.99872	0.38877	0.48796	0.48406
22	SH2B3	12	110323135	110378810	0.69064	0.46967	0.36973	0.62359
23	rs10850411	12	113822179	113922179	0.99999	0.96044	0.92497	0.91949
24	FES	15	89222929	89245010	0.96616	0.26903	0.35591	0.55000
25	rs1378942	15	72814420	72914420	0.28300	0.02224	0.02575	0.02478
26	GOSR2	17	42350482	42465002	0.78478	0.49338	0.45838	0.57622
27	ZNF652	17	44716567	44799834	0.98444	0.63856	0.76786	0.88685
28	rs1327235	20	10867030	10967030	0.92765	0.48760	0.48532	0.55831
29	rs6015450	20	57134512	57234512	0.40123	0.24997	0.25722	0.17082

Table B.21: Meta-Analysis of dBP (6361 subjects) in MESA. Four ethnic groups were combined using Fisher's method.

	name	chr	start	end	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	0.65418	0.61873	0.56224	0.72836
2	MTHFR	1	11763367	11794564	0.13322	0.00934	0.00557	0.00854
3	rs13082711	3	27462913	27562913	0.00434	0.00041	0.00063	0.00241
4	MECOM	3	170278981	170869100	0.40581	0.04256	0.10761	0.07974
5	ULK4	3	41258094	41983926	0.12987	0.55612	0.56460	0.87757
6	SLC39A8	4	103386221	103576438	0.97592	0.87205	0.69878	0.34243
7	GUCY1A3	4	156802313	156877951	0.82750	0.19607	0.24412	0.47640
8	rs1458038	4	81333747	81433747	0.99279	0.32604	0.26011	0.26061
9	rs1173771	5	32800785	32900785	0.50600	0.82381	0.76341	0.87893
10	rs11953630	5	157727980	157827980	0.90491	0.48791	0.58455	0.46902
11	HFE	6	26190488	26211550	0.78654	0.18630	0.31383	0.39228
12	rs805303	6	31674345	31774345	0.11321	0.04401	0.04039	0.05240
13	rs4373814	10	18409978	18509978	0.86003	0.62699	0.50342	0.67279
14	PLCE1	10	95738736	96083139	1.00000	0.90959	0.90494	0.82606
15	CACNB2	10	18464612	18875804	0.23328	0.07493	0.05311	0.04745
16	C10orf107	10	63087725	63201530	0.02302	0.00146	0.00097	0.00086
17	NT5C2	10	104830930	104948046	1.00000	0.77595	0.68184	0.76498
18	rs7129220	11	10257114	10357114	0.47707	0.08902	0.09892	0.18138
19	ARHGAP42	11	100058594	100371866	0.63454	0.44668	0.29560	0.56021
20	PLEKHA7	11	16751418	16997566	0.70240	0.18356	0.13426	0.11532
21	ATP2B1	12	88500959	88632208	0.50290	0.25040	0.29796	0.15524
22	SH2B3	12	110323135	110378810	0.26183	0.71943	0.63767	0.55636
23	rs10850411	12	113822179	113922179	0.95580	0.90671	0.88300	0.89773
24	FES	15	89222929	89245010	1.00000	0.56643	0.57727	0.63884
25	rs1378942	15	72814420	72914420	0.62576	0.23689	0.16497	0.16428
26	GOSR2	17	42350482	42465002	0.58833	0.32515	0.41807	0.53758
27	ZNF652	17	44716567	44799834	0.93545	0.65153	0.70855	0.76768
28	rs1327235	20	10867030	10967030	1.00000	0.86045	0.92321	0.97171
29	rs6015450	20	57134512	57234512	0.64413	0.20782	0.23223	0.28420

APPENDIX C

Supplementary Materials for Chapter IV

1. Detailed Proofs

1.1 Proof of Result 3.1

To derive the asymptotic distribution of $Q = \frac{1}{m} S^T_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$, the first step is to show $\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$ asymptotically follows a multivariate normal distribution with mean zero.

$$\begin{aligned}
\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) &= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i) \\
&= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) - \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)
\end{aligned}$$

By Taylor expansion at $\boldsymbol{\zeta}$, $\widehat{\boldsymbol{\zeta}}$ can be replaced by $\boldsymbol{\zeta}$ except an $o_p(1)$ term. It is easy to show the score vector

$$\begin{aligned}
\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) &= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \qquad \text{(C.1)} \\
&- \{\mathbb{E}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)] + o_p(1)\}\sqrt{m}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1). \qquad \text{(C.2)}
\end{aligned}$$

Then we deal with the estimation of $\boldsymbol{\beta}$. We note $\widehat{\boldsymbol{\beta}}$ is the solution of an estimating equation

$$\frac{1}{\sqrt{m}} S_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i) = 0.$$

By the Taylor expansion at $(\boldsymbol{\beta}, \boldsymbol{\zeta})$, the convergence of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}})$ and the strong law of large numbers, we have

$$\sqrt{m}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \{\mathbb{E}[(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)] + o_p(1)\}^{-1}[\frac{1}{\sqrt{m}} S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0) + o_p(1)].$$

We plug $\sqrt{m}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ into the equation (C.2), then the score vector

$$\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) = [\boldsymbol{A} + o_p(1)] \frac{1}{\sqrt{m}} [S_{\boldsymbol{\gamma}}^T(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0), S_{\boldsymbol{\beta}}^T(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)]^T + o_p(1),$$

where $\boldsymbol{A} = \{\boldsymbol{I}_q, -\mathbb{E}[(\boldsymbol{E}_i {*} \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)]\mathbb{E}^{-1}[(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)^T \boldsymbol{V}^{-1}(\boldsymbol{\zeta})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)]\}$.

By the central limit theorem and the unbiasedness of score functions $S_{\boldsymbol{\gamma}}$ and $S_{\boldsymbol{\beta}}$,

$$\frac{1}{\sqrt{m}} [S_{\boldsymbol{\gamma}}^T(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0), S_{\boldsymbol{\beta}}^T(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)]^T \Rightarrow N(0, \boldsymbol{D}),$$

where $\boldsymbol{D} = var[\frac{1}{\sqrt{m}} S(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)]$. Thus,

$$\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) = N(0, \boldsymbol{\Sigma}) + o_p(1)$$

where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T$.

So far, we have proved $\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$ asymptotically follows a multivariate normal distribution with mean zero and covariance $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T$. By the property of quadratic forms, $Q = \frac{1}{m} S_{\boldsymbol{\gamma}}^T(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)$ is asymptotically distributed as

$$\sum_{k=1}^{q} \lambda_k \chi_k^2 \tag{C.3}$$

where χ_k^2s are i.i.d. Chi-square distributions with degree of freedom one; $\lambda_1 \geq \ldots \geq \lambda_q$ are the ordered eigen-values of $\boldsymbol{\Sigma}$. Next, we estimate $\boldsymbol{A}$ and $\boldsymbol{D}$ by

$$\widehat{\boldsymbol{A}} = \{\boldsymbol{I}_q, -[\sum_{i=1}^{m} (\boldsymbol{E}_i {*} \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)][\sum_{i=1}^{m} (\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)]^{-1}\},$$

$$\widehat{\boldsymbol{D}} = \frac{1}{m-p-q-1} \sum_{i=1}^{m} S_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) S_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)^T, S_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) = [S_{\boldsymbol{\gamma},i}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)^T, S_{\boldsymbol{\beta},i}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0)^T]^T,$$

and show that the weights in equation (C.3) can be estimated consistently by the eigenvalues of $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{A}}^T$. First we show $\widehat{\boldsymbol{A}} \rightarrow \boldsymbol{A}$. We note

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{E}_i * \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) &= \frac{1}{m}\sum_{i}(\boldsymbol{E}_i * \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) \\ &+ (\zeta_l^* - \zeta_l)\frac{1}{m}\sum_{i=1}^{m}\sum_{l}(\boldsymbol{E}_i * \boldsymbol{G}_i)^T\frac{\partial \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})}{\partial \zeta_l}(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i).\end{aligned}$$

By the strong law of large numbers,

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{E}_i * \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) = \frac{1}{m}\sum_{i}(\boldsymbol{E}_i * \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) + o_p(1).$$

Similar argument can be applied to $\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)$. Therefore,

$$\begin{aligned}\widehat{\boldsymbol{A}} &= \{\boldsymbol{I}_q, -[\sum_{i=1}^{m}(\boldsymbol{E}_i * \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)][\sum_{i=1}^{m}(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)]^{-1}\} \\ &= \{\boldsymbol{I}_q, -[\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{E}_i * \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)][\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)]^{-1}\} \\ &= \boldsymbol{A} + o_p(1)\end{aligned}$$

Then we show $\widehat{\boldsymbol{D}} \rightarrow \boldsymbol{D}$. It suffices to show $\sup_{\|\boldsymbol{c}\|=1} |\boldsymbol{c}^T(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{c}| = o_p(1)$ where $\boldsymbol{c}$ is an arbitrary vector with norm one; $\boldsymbol{D} = var[\frac{1}{\sqrt{m}}S(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)]$ and $\widehat{\boldsymbol{D}}$ is defined as before. Let

$$\begin{aligned}\widehat{\boldsymbol{D}} &= \frac{1}{m}\sum_{i}\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})\boldsymbol{B}_i, \\ \widetilde{\boldsymbol{D}} &= \frac{1}{m}\sum_{i}\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{B}_i, \\ \boldsymbol{D} &= \frac{1}{m}\sum_{i}\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})var(\boldsymbol{\varepsilon}_i)\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{B}_i,\end{aligned}$$

where $\boldsymbol{B}_i = (\boldsymbol{E}_i * \boldsymbol{G}_i, \boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)$, $\boldsymbol{\varepsilon}_i = \boldsymbol{Y}_i - \boldsymbol{\mu}_i$ and $\widehat{\boldsymbol{\varepsilon}}_i = \boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i$. We have

$$|\boldsymbol{c}^T(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{c}| \leq |\boldsymbol{c}^T(\widehat{\boldsymbol{D}} - \widetilde{\boldsymbol{D}})\boldsymbol{c}| + |\boldsymbol{c}^T(\widetilde{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{c}|.$$

$$\begin{aligned}
|\boldsymbol{c}^T(\widehat{\boldsymbol{D}}-\widetilde{\boldsymbol{D}})\boldsymbol{c}| &\leq |\frac{1}{m}\sum_i \boldsymbol{c}^T\boldsymbol{B}_i^T[\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})-\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})]\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})\boldsymbol{B}_i\boldsymbol{c}| \\
&+ |\frac{1}{m}\sum_i \boldsymbol{c}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i^T[\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})-\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})]\boldsymbol{B}_i\boldsymbol{c}| \\
&+ |\frac{1}{m}\sum_i \boldsymbol{c}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i^T-\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^T)\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{B}_i\boldsymbol{c}| \\
&\triangleq J_{n1}+J_{n2}+J_{n3},
\end{aligned}$$

where

$$\begin{aligned}
\sup_{\|\boldsymbol{c}\|=1} J_{n1} &\leq \sup_{\|\boldsymbol{c}\|=1}\frac{1}{m}\sum_i \|B_i^T\boldsymbol{c}\|^2\|\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})\widehat{\boldsymbol{\varepsilon}}_i\|\|[\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})-\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})]\widehat{\boldsymbol{\varepsilon}}_i\| \\
&\leq \sup_{\|\boldsymbol{c}\|=1}\frac{1}{m}\sum_i \|B_i^T\boldsymbol{c}\|^2 C o_p(1) \leq \lambda_{max}(\frac{1}{m}\boldsymbol{B}^T\boldsymbol{B})C o_p(1) = o_p(1).
\end{aligned}$$

Similarly, $\sup_{\|\boldsymbol{c}\|=1} J_{n2} = o_p(1)$. With respect to the third term J_{n3},

$$\begin{aligned}
\sup_{\|\boldsymbol{c}\|=1} J_{n3} &\leq \sup_{\|\boldsymbol{c}\|=1}|\frac{1}{m}\sum_i \boldsymbol{c}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\widehat{\boldsymbol{\varepsilon}}_i(\widehat{\boldsymbol{\varepsilon}}_i-\boldsymbol{\varepsilon}_i)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{B}_i\boldsymbol{c}| \\
&+ \sup_{\|\boldsymbol{c}\|=1}|\frac{1}{m}\sum_i \boldsymbol{c}^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\varepsilon}}_i-\boldsymbol{\varepsilon}_i)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{B}_i\boldsymbol{c}| \\
&\leq \sup_{\|\boldsymbol{c}\|=1}\frac{1}{m}\sum_i \|\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\|^2(\|\boldsymbol{\varepsilon}_i\|+\|\widehat{\boldsymbol{\varepsilon}}_i\|)\|\widehat{\boldsymbol{\varepsilon}}_i-\boldsymbol{\varepsilon}_i\|\|\boldsymbol{B}_i\boldsymbol{c}\| \\
&\leq C\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\sup_{\|\boldsymbol{c}\|=1}\frac{1}{m}\sum_i \|\boldsymbol{B}_i\boldsymbol{c}\| = C\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\lambda_{max}(\frac{1}{m}\boldsymbol{B}^T\boldsymbol{B}) = o_p(1).
\end{aligned}$$

In addition, $|\boldsymbol{c}^T(\widetilde{\boldsymbol{D}}-\boldsymbol{D})\boldsymbol{c}| = o_p(1)$ by the strong law of large numbers. Thus,

$$\sup_{\|\boldsymbol{c}\|=1}|\boldsymbol{c}^T(\widehat{\boldsymbol{D}}-\boldsymbol{D})\boldsymbol{c}| = o_p(1).$$

The final step is to show the weights in equation (C.3) can be estimated consistently by the eigen-values of $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{A}}^T$. Since $\widehat{\boldsymbol{A}} \to \boldsymbol{A}$ and $\widehat{\boldsymbol{D}} \to \boldsymbol{D}$,

$$\max\{|\lambda_{min}(\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{A}}^T-\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T)|, |\lambda_{max}(\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{A}}^T-\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T)|\} = o_p(1).$$

By Weyl's inequality (Franklin, 1993),

$$\lambda_{min}(\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{A}}^T-\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T) \leq |\widehat{\lambda}_k-\lambda_k| \leq \lambda_{max}(\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{A}}^T-\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T)$$

Finally,

$$\max_{1\le k\le q} |\widehat{\lambda}_k - \lambda_k| = o_p(1), \quad n \to \infty.$$

1.2 Proof of Result 4.1

Let $v_{ij,ik}$ be the (j,k)-th element of $\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})$. We let

$$E_{ij} = E_{ij}^* - \frac{\sum_{i,j} E_{ij}^* c_{ij}}{\sum_{i,j} c_{ij}}, \quad G_i = G_i^* - \frac{1}{m}\sum_i G_i^*$$

where $c_{ij} = \sum_k v_{ij,ik}$; E_{ij}^* and G_i^* are the original exposure and genetic variant. In practice, $\boldsymbol{\zeta}$ is replaced by its estimator $\widehat{\boldsymbol{\zeta}}$.

Since $S_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)$ is a vector where each element corresponds to one genetic variable, we prove the result when there is only one genetic variable. Suppose the true main effects are $h_E(\boldsymbol{E}_i)$ and $h_G(\boldsymbol{G}_i)$,

$$\begin{aligned}
\mathbb{E}_{H_0}[S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)] &= \mathbb{E}_{H_0}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{X}_i^E \boldsymbol{\beta}_{X^E}^0 - \boldsymbol{X}_i^E \boldsymbol{\beta}_{X^E})] \\
&+ \mathbb{E}_{H_0}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{X}_i^G \boldsymbol{\beta}_{X^G}^0 - \boldsymbol{X}_i^G \boldsymbol{\beta}_{X^G})] \\
&+ \mathbb{E}_{H_0}\{(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[h_E(\boldsymbol{E}_i) - \boldsymbol{E}_i \beta_E]\} \\
&+ \mathbb{E}_{H_0}\{(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[h_G(\boldsymbol{G}_i) - \boldsymbol{G}_i \beta_G]\} \\
&\triangleq K_{X^E} + K_{X^G} + K_E + K_G,
\end{aligned}$$

where $\boldsymbol{\beta}_{X^E}^0$ and $\boldsymbol{\beta}_{X^E}^0$ are the coefficients in the true model. We first note

$$\begin{aligned}
K_{X^E} &= \mathbb{E}_{H_0}[\sum_{(j,k)} E_{ij} G_i v_{ij,ik}(\boldsymbol{X}_{i,k}^E \boldsymbol{\beta}_{X^E}^0 - \boldsymbol{X}_{i,k}^E \boldsymbol{\beta}_{X^E})] \\
&= \mathbb{E}_{H_0}[\sum_{(j,k)} E_{ij} v_{ij,ik}(\boldsymbol{X}_{i,k}^E \boldsymbol{\beta}_{X^E}^0 - \boldsymbol{X}_{i,k}^E \boldsymbol{\beta}_{X^E})]\mathbb{E}_{H_0}(G_i) = 0,
\end{aligned}$$

where the second equality is due to the independence between $\boldsymbol{G}_i$ and $(\boldsymbol{X}_i^E, \boldsymbol{E}_i)$; the last equality is because G_i is centered. With respect to the second term K_{X^G}.

$$\begin{aligned} K_{X^G} &= \mathbb{E}_{H_0}[\sum_{(j,k)} E_{ij} G_i v_{ij,ik} (\boldsymbol{X}^G_{i,k}\boldsymbol{\beta}^0_{X^G} - \boldsymbol{X}^G_{i,k}\boldsymbol{\beta}_{X^G})] \\ &= (\boldsymbol{\beta}^0_{X^G} - \boldsymbol{\beta}_{X^G}) \sum_{(j,k)} \mathbb{E}_{H_0}(E_{ij} v_{ij,ik}) cov(G_i, \boldsymbol{X}^G_{i,k}) \\ &= \mathbb{E}_{H_0}(\sum_j E_{ij} \sum_k v_{ij,ik}) c_i = 0, \end{aligned}$$

where $c_i = cov(G_i, \boldsymbol{X}^G_{i,k})(\boldsymbol{\beta}^0_{X^G} - \boldsymbol{\beta}_{X^G})$ is a constant which depends on neither j nor k, because $cov(\boldsymbol{X}^G_{i,k}, G_i)$ is time invariant. The weights used to center E ensures the last equality. Similar argument can be applied to K_E and K_G,

$$\begin{aligned} K_E &= \mathbb{E}_{H_0}\{\sum_{(j,k)} E_{ij} G_i v_{ij,ik} [h_E(E_{i,k}) - E_{i,k}\beta_E]\} \\ &= \mathbb{E}_{H_0}\{\sum_{(j,k)} E_{ij} v_{ij,ik} [h_E(E_{i,k}) - E_{i,k}\beta_E]\} \mathbb{E}_{H_0}(G_i) = 0. \end{aligned}$$

$$\begin{aligned} K_G &= \mathbb{E}_{H_0}\{\sum_{(j,k)} E_{ij} G_i v_{ij,ik} [h_G(G_i) - G_i\beta_G]\} \\ &= \mathbb{E}_{H_0}(\sum_j E_{ij} \sum_k v_{ij,ik}) \mathbb{E}_{H_0}\{G_i[h_G(G_i) - G_i\beta_G]\} = 0, \end{aligned}$$

Finally,

$$\mathbb{E}_{H_0}[S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)] = 0.$$

The above derivation is for balanced data. The proof follows similarly if the data is missing completely at random.

1.3 Proof of Result 4.2

Since $S_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)$ is a vector where each element separately corresponds to one genetic variable, we prove the result when there is only one genetic variable without loss of generality.

$$\begin{aligned} \frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) &= \frac{1}{\sqrt{m}} \sum_i (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i) \\ &= \frac{1}{\sqrt{m}} \sum_i (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^0) - \frac{1}{\sqrt{m}} \sum_i (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^0) \end{aligned}$$

By Taylor expansion, we can replace $\widehat{\boldsymbol{\zeta}}$ by $\boldsymbol{\zeta}$ except an $o_p(1)$ term. Thus

$$\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) = \frac{1}{\sqrt{m}} \sum_i (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^0) - \frac{1}{\sqrt{m}} \sum_i (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^0) + o_p(1).$$

We will show the second term $\frac{1}{\sqrt{m}} \sum_i (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\zeta}})(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^0) = o_p(1)$. This is by evaluating the difference between the estimated model and true model.

$$\begin{aligned}
\frac{1}{\sqrt{m}} \sum_i (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^0) &= (\widehat{\boldsymbol{\beta}}_{X^E} - \boldsymbol{\beta}_{X^E}^0) \frac{1}{\sqrt{m}} \sum_{i=1}^m (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}) \boldsymbol{X}_i^E \\
&+ (\widehat{\boldsymbol{\beta}}_{X^G} - \boldsymbol{\beta}_{X^G}^0) \frac{1}{\sqrt{m}} \sum_{i=1}^m (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}) \boldsymbol{X}_i^G \\
&+ \frac{1}{\sqrt{m}} \sum_{i=1}^m (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}) [h_{E,U}(\boldsymbol{E}_i, \widehat{\boldsymbol{\beta}}_E) - h_E(\boldsymbol{E}_i)] \\
&+ (\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}_G^0) \frac{1}{\sqrt{m}} \sum_{i=1}^m (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}) \boldsymbol{G}_i \\
&\triangleq I_{m,X^E} + I_{m,X^G} + I_{m,E} + I_{m,G}
\end{aligned}$$

We note that $h_{E,U}(x; \widehat{\boldsymbol{\beta}})$ uniformly converges to $h_E(x)$ for $\forall\, x$ as $m \to \infty$ and $(\widehat{\boldsymbol{\beta}}_{X^E}^T, \widehat{\boldsymbol{\beta}}_{X^G}^T, \widehat{\boldsymbol{\beta}}_G^T)^T$ converge to $(\boldsymbol{\beta}_{X^E}^{0T}, \boldsymbol{\beta}_{X^G}^{0T}, \boldsymbol{\beta}_G^{0T})^T$. To evaluate $I_{m,X^E} = (\widehat{\boldsymbol{\beta}}_{X^E} - \boldsymbol{\beta}_{X^E}^0) \frac{1}{\sqrt{m}} \sum_{i=1}^m (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}) \boldsymbol{X}_i^E$, by C1) and C2), following the similar argument in the proof of Result 4.1,

$$E[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}^*) \boldsymbol{X}_i^E] = E(\sum_{j,k} E_{i,j} G_i v_{ij,ik} \boldsymbol{X}_{i,k}^E) = E(\sum_{j,k} E_{i,j} v_{ij,ik} \boldsymbol{X}_{i,k}^E) E(G_i) = 0.$$

By the central limit theorem and convergence of $\widehat{\boldsymbol{\beta}}_{X^E}$, we have

$$I_{m,X^E} = (\widehat{\boldsymbol{\beta}}_{X^E} - \boldsymbol{\beta}_{X^E}^0) \frac{1}{\sqrt{m}} \sum_i (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}^*) \boldsymbol{X}_i^E = o_p(1)$$

where the weights used to center E ensures that $(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}^*) \boldsymbol{X}_i^E$ has mean zero. Similarly $I_{m,X^G} = I_{m,G} = o_p(1)$. To evaluate $I_{m,E}$, we note

$$\begin{aligned}
I_{m,E} &= \frac{1}{\sqrt{m}} \sum_{i=1}^m (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}) [\hat{h}_{E,U}(\boldsymbol{E}_i) - h_{E,U}(\boldsymbol{E}_i)] \\
&= o_p(1) \frac{1}{\sqrt{m}} \sum_{i=1}^m (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta}) \mathbf{1} = \mathbf{o_p(1)},
\end{aligned}$$

where the last equality is by then central limit theorem. Finally,

$$\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}, 0) = \frac{1}{\sqrt{m}} \sum_i (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[\boldsymbol{Y}_i - \mathbb{E}^0_{H_0}(\boldsymbol{Y}_i | \boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)] + o_p(1)$$

The above derivation is for balanced data. The proof follows similarly if the data is missing completely at random.

1.4 Proof of Result 4.3

We prove the result by contrasting the true model where all PCs are included and a reduced model with the S PCs only:

$$\mu^0_{i,j} = \boldsymbol{X}^T_{i,j}\boldsymbol{\beta}^0_X + E_{i,j}\beta^0_E + \sum_{s=1}^{q} P^s_i \beta^0_{P,s} \quad \mu_{i,j} = \boldsymbol{X}^T_{i,j}\boldsymbol{\beta}_X + E_{i,j}\beta_E + \sum_{s=1}^{S} P^s_i \beta_{P,s}.$$

We note that $\boldsymbol{\beta} \triangleq (\boldsymbol{\beta}^T_X, \beta_E, \beta_{P,1}, ..., \beta_{P,S})^T$ is the solution of

$$\mathbb{E}_{H_0}[(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}^1_i, \ldots, \boldsymbol{P}^S_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)] = 0.$$

By calculating the expectation and some algebra, we have

$$\text{(C.4)} \quad (\boldsymbol{\beta}^T_X, \beta_E, \beta_{P,1}, ..., \beta_{P,S})^T = (\boldsymbol{\beta}^{0T}_X, \beta^0_E, \beta^0_{P,1}, ..., \beta^0_{P,S})^T + \boldsymbol{A}^{-1}\boldsymbol{B}(\beta^0_{P,S+1}, ..., \beta^0_{P,q})^T,$$

where

$$\begin{aligned} \boldsymbol{A} &= \mathbb{E}_{H_0}\{[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}^1_i, \ldots, \boldsymbol{P}^S_i]^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}^1_i, \ldots, \boldsymbol{P}^S_i]\} \\ \boldsymbol{B} &= \mathbb{E}_{H_0}\{[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}^1_i, \ldots, \boldsymbol{P}^S_i]^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[\boldsymbol{P}^{S+1}_i, \ldots, \boldsymbol{P}^q_i]\}. \end{aligned}$$

Then we will evaluate the bias due to fitting the reduced model,

$$\mathbb{E}_{H_0}[S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)] = \mathbb{E}_{H_0}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)] = \mathbb{E}_{H_0}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{\mu}^0_i - \boldsymbol{\mu}_i)].$$

By plugging (C.4) into the above equation,

$$\begin{aligned}
\mathbb{E}_{H_0}[S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta},\boldsymbol{\zeta},0)] &= -\mathbb{E}_{H_0}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S)]\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{\beta}_{P,S+1}^0, \ldots, \boldsymbol{\beta}_{P,q}^0)^T \\
&+ \sum_{s=S+1}^{q} \mathbb{E}_{H_0}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{P}_i^s]\beta_{P,s}^0 \\
&= -\sum_{s=S+1}^{q} \mathbb{E}_{H_0}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S)]\boldsymbol{A}^{-1}\boldsymbol{b}^s\beta_{P,s}^0 \\
&+ \sum_{s=S+1}^{q} \mathbb{E}_{H_0}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{P}_i^s]\beta_{P,s}^0 \\
&= \sum_{s=S+1}^{q} \{\mathbb{E}_{H_0}[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{P}_i^s] - \boldsymbol{\phi}^s\}\beta_{P,s}^0,
\end{aligned}$$

where $\boldsymbol{b}^s$ is the s-th column of $\boldsymbol{B}$; $\boldsymbol{\phi}^s = E\{(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S]\}\boldsymbol{A}^{-1}\boldsymbol{b}^s$.

2. Additional Numerical Studies

2.1 Main effect adjustment of G when the number of SNP exceeds the sample size

Table C.1: Simulation study evaluating the main effect adjustment of G (500 subjects). A linear main effect of E was fitted. Each cell presents the type I error rate/power based on 1000 replicates. MinP: single SNP analysis using GEE adjusted by the effective number of independent tests (Gao et al., 2008). iSKAT: region based test proposed by Lin et al. (2013). rareGE: rareGE test proposed by Chen et al. (2014) assuming a random main effect of G. GE-none: the proposed test adjusting for none of the SNPs. GE-wPCA/wPCAM-$\sqrt{m}$: the proposed test adjusting for the leading $\sqrt{m}$ components using the weighted PCA and robust(wPCA)/model-based(wPCAM) inference. GE-true: the proposed test with the correct model, which correctly correctly includes all SNPs with non-zero main effects. Type I error rate and power were both evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07.

Cross-sectional data								
	Type I error rate				Power			
	C1 holds		C1 does not hold		C1 holds		C1 does not hold	
q	400	700	400	700	400	700	400	700
MinP	0.052	0.047	0.116	0.129	0.426	0.355	0.676*	0.636*
iSKAT	0.083	1.000	0.085	1.000	0.334*	0*	0.561*	0*
rareGE	0.049	0.061	0.062	0.072	0.433	0.411	0.690	0.616*
GE-none	0.030	0.021	0.093	0.101	0.281	0.256	0.623	0.549
GE-wPCA -$\sqrt{m}$	0.032	0.013	0.031	0.043	0.343	0.317	0.574	0.526
GE-wPCAM-$\sqrt{m}$	0.042	0.046	0.029	0.048	0.423	0.402	0.638	0.607
GE-true	0.031	0.028	0.040	0.033	0.352	0.325	0.610	0.597

Longitudinal data								
	Type I error rate				Power			
	Ind. holds		Ind. does not hold		Ind. holds		Ind. does not hold	
q	400	700	400	700	400	700	400	700
MinP	0.044	0.029	0.084	0.069	0.325	0.298	0.392*	0.438
GE-none	0.035	0.043	0.110	0.126	0.381	0.381	0.368*	0.396*
GE-wPCA -$\sqrt{m}$	0.024	0.034	0.025	0.028	0.402	0.380	0.520	0.520
GE-wPCAM-$\sqrt{m}$	0.022	0.043	0.030	0.040	0.438	0.435	0.573	0.581
GE-true	0.030	0.040	0.026	0.033	0.431	0.426	0.563	0.554

2.2 Detailed evaluation of PCA, PLS and weighted PCA

Table C.2: Simulation study evaluating the main effect adjustment of G (775 subjects). A linear main effect of E was included in the working model. Each cell presents the type I error rate/power based on 1000 replicates. GE-PCA/PLS/wPCA-5/$\sqrt{m}$: the proposed test adjusting for the leading 5/$\sqrt{n}$ components using principal component analysis (PCA)/partial least square regression (PLS)/ the weighted PCA (wPCA) approach. GE-true: the proposed test with a correct model. Type I error rate and power were both evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has high type I error rate (> 0.07). The zero calibrated power is due to extremely low p-values when type I error rate is high.

Cross-sectional data								
	Type I error rate				Power			
	Ind. holds		Ind. does not hold		Ind. holds		Ind. does not hold	
q	400	700	400	700	400	700	400	700
GE-PCA-5	0.036	0.038	0.122	0.107	0.517	0.444	0.706*	0.682*
GE-PLS-5	0.037	0.035	0.044	0.035	0.547	0.476	0.747	0.718
GE-wPCA-5	0.038	0.042	0.072	0.076	0.540	0.480	0.738*	0.708*
GE-PCA-$\sqrt{m}$	0.039	0.040	0.052	0.041	0.560	0.481	0.785	0.744
GE-PLS-$\sqrt{m}$	0.042	0.032	0.033	0.029	0.499	0.381	0.714	0.642
GE-wPCA -$\sqrt{m}$	0.038	0.038	0.044	0.041	0.557	0.483	0.766	0.742
GE-true	0.036	0.038	0.045	0.041	0.584	0.509	0.797	0.759

Longitudinal data								
	Type I error rate				Power			
	Ind. holds		Ind. does not hold		Ind. holds		Ind. does not hold	
q	400	700	400	700	400	700	400	700
GE-PCA-5	0.043	0.033	0.086	0.094	0.595	0.578	0.707	0.714
GE-PLS-5	0.039	0.028	0.035	0.033	0.603	0.586	0.702	0.687
GE-wPCA-5	0.044	0.036	0.062	0.057	0.599	0.577	0.705	0.703
GE-PCA-$\sqrt{m}$	0.037	0.025	0.031	0.040	0.588	0.559	0.696	0.688
GE-PLS-$\sqrt{m}$	0.030	0.049	0.031	0.033	0.571	0.566	0.662	0.633
GE-wPCA -$\sqrt{m}$	0.033	0.034	0.031	0.034	0.588	0.569	0.687	0.682
GE-true	0.040	0.034	0.038	0.039	0.615	0.589	0.724	0.707

3. Additional Data Analysis

3.1 Detailed description of the neighborhood features

The four neighborhood measures include two geographic information system based measures and two survey based measures: 1. Density of favorable food stores (GIS-based); 2. Density of recreational facilities (GIS-based); 3. Perceived healthy foods availability (survey-based); 4. Perceived walkability (survey-based). The GIS measures are from the National Establishment Time Series (NETS) database from Wall and Associates for 2000 to 2007, where data on food stores and commercially-available recreational facilities for every ZIP code within a 5 miles radius of MESA participant households were collected. The food stores were identified from a total of 15 Standardized Industrial Codes (SIC) and the data was enhanced by adding supermarket data from Nielsen (2008)/TDLinx as in Auchincloss et al. (2012). The recreational facilities were identified from 114 SICs and includes indoor conditioning, dance, bowling, golf, team and racquet sports, and water activities. Gaussian kernel weighted densities of the food stores and recreational facilities were calculated for one mile buffers surrounding participant households using ArcGIS 9.3 for each year of the MESA examination. The survey based measures of healthy food availability and walkability were obtained from questionnaires administered to MESA participants. Respondents were asked to answer a set of questions regarding healthy food availability (large selections of fresh fruit and vegetables and low fat foods) and walkability (pleasurability and ease of walking, and frequency of other people walking or exercising in the neighborhood) using a 5-point scale. The measures were then aggregated by pooling all available respondents in each census tract using Conditional Empirical Bayes estimates adjusted for respondent age, sex, source and site as described in Mujahid et al. (2008). For all these measures, higher values indicate a better neighborhood environment.

3.2 Linkage disequilibrium structure of region MECOM

Figure C.1: Region MECOM (821 SNPs) in Chinese Americans (775 subjects). The top figure shows its linkage disequilibrium structure, created by Haploview based on HapHap CHD reference samples. Bar plots show the distribution of coefficient squares and eigen-values. The coefficients were estimated by first regressing systolic blood pressure on covariates used in the data analysis of MESA, then regress the residuals on each SNP/PC.

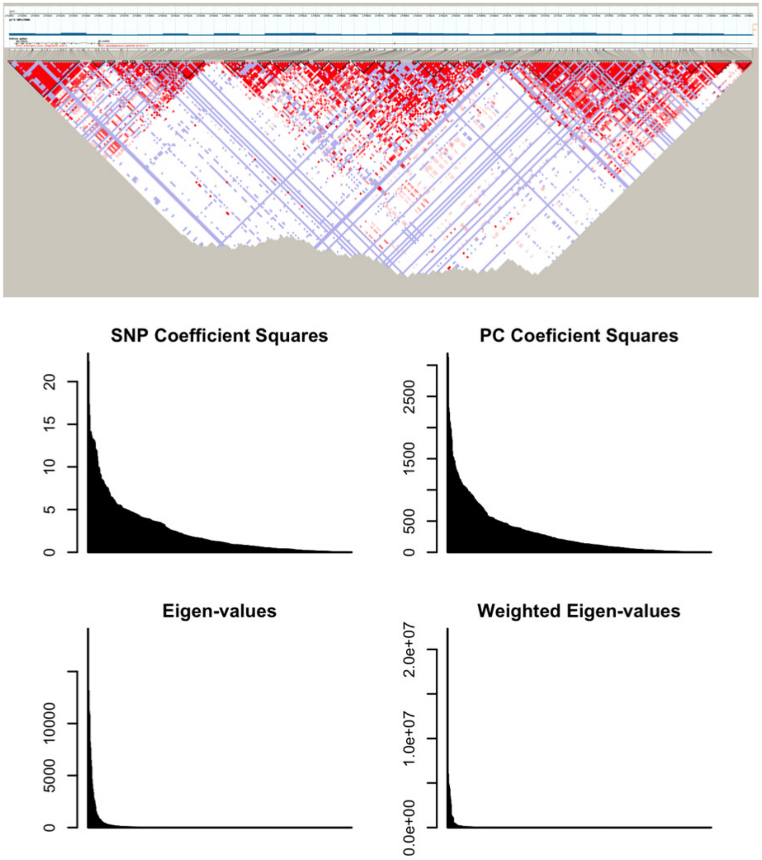


Table C.3: Gender distribution of MESA subjects across site and race. Each cell represents the number of subject in the corresponding category. WFU: Wake Forest University, Winston Salem, NC; COL: Columbia University, New York, NY; JHU: Johns Hopkins University, Baltimore, MD; UMN: University of Minnesota, Twin Cities, MN; NWU: Northwestern University, Chicago, IL; UCLA: University of California - Lost Angeles, Los Angeles, CA.

	Gender		
Site	Female	Male	All
WFU	528	464	992
COL	536	434	970
JHU	556	488	1044
UMN	532	518	1050
NWU	551	508	1059
UCLA	666	648	1314
All	3369	3060	6429

	Gender		
Race	Female	Male	All
White/Caucasian	1321	1206	2527
Chinese American	394	381	775
Black/African-American	906	771	1677
Hispanic	748	702	1450
All	3369	3060	6429

Table C.4: Marginal association between sBP/dBP and the neighborhood exposures. This is a joint analysis of all four ethnic groups using GEE, adjusting for age, gender, BMI and the socioeconomic status.

	sBP			dBP		
Exposure	Coefficient	Sd	P-value	Coefficient	Sd	P-value
Density of Favorable Food Stores	-0.15	0.05	3.65E-03	0.03	0.03	1.83E-01
Density of Recreational Facilities	-0.08	0.02	4.69E-04	6.68E-05	0.01	9.95E-01
Perceived Healthy Foods Availability	-2.18	0.44	8.74E-07	-0.61	0.21	4.21E-03
Perceived Walkability	0.05	0.81	9.48E-01	0.58	0.38	1.27E-01

Table C.5: Longitudinal summary of blood pressure phenotypes and covariates we adjusted for in MESA across four exams. Sd: standard deviation. N: number of subject. sBP: systolic blood pressure. dBP: diastolic blood pressure. BMI: body mass index. SES: socioeconomic status. DFFS: density of favorable food stores. DRF: density of recreational facilities. PHFA: perceived healthy foods availability. PW: perceived walkability.

	Exam 1 (24 months)			Exam 2 (18 months)			Exam 3 (18 months)			Exam 4 (24 months)		
	Mean	Sd	N	Mean	Sd	N	Mean	Sd	N	Mean	Sd	N
sBP (mm Hg)	126.51	21.55	6427	124.33	20.79	5898	123.16	20.58	5619	123.59	20.56	5399
dBP (mm Hg)	71.82	10.27	6427	70.37	10.09	5898	69.69	9.94	5619	69.61	10.05	5399
BMI (kg/m^2)	28.3	5.47	6429	28.33	5.48	5889	28.28	5.51	5621	28.38	5.58	5402
Age (years)	62.22	10.24	6429	63.69	10.1	5900	64.99	9.99	5628	66.51	9.94	5505
SES	-0.24	1.36	5853	-0.29	1.35	5755	-0.45	1.33	5576	-0.74	1.32	5333
DFFS	2.59	4.22	5853	2.64	4.32	5755	2.6	4.34	5595	2.66	4.45	5383
DRF	4.71	8.29	5853	5.37	9.38	5755	5.73	9.76	5595	6.61	11.32	5383
PHFA	3.48	0.48	5785	3.49	0.47	5659	3.48	0.47	5398	3.72	0.38	4932
PW	3.91	0.31	5785	3.91	0.3	5659	3.91	0.3	5398	3.95	0.26	4932

Table C.6: Chromosomal Region Information for the 29 regions considered in the MESA analysis.

Region Name	Chromosome	Start	End	Index SNP	Nearest Gene	Coded Allele Frequency
MOV10	1	113012286	113049891	rs2932538	MOV10	0.75
rs13082711	3	27462913	27562913	rs13082711	SLC4A7	0.78
MECOM	3	170278981	170869100	rs419076	MECOM	0.47
SLC39A8	4	103386221	103576438	rs13107325	SLC39A8	0.05
GUCY1A3	4	156802313	156877951	rs13139571	GUCY1A3,GUCY1B3	0.76
rs1173771	5	32800785	32900785	rs1173771	NPR3,C5orf23	0.6
rs11953630	5	157727980	157827980	rs11953630	EBF1	0.37
HFE	6	26190488	26211550	rs1799945	HFE	0.14
rs805303	6	31674345	31774345	rs805303	BAT2,BAT5	0.61
rs4373814	10	18409978	18509978	rs4373814	CACNB2	0.55
PLCE1	10	95738736	96083139	rs932764	PLCE1	0.44
rs7129220	11	10257114	10357114	rs7129220	ADM	0.89
ARHGAP42	11	100058594	100371866	rs633185	FLJ32810,TMEM133	0.28
FES	15	89222929	89245010	rs2521501	FURIN,FES	0.31
GOSR2	17	42350482	42465002	rs17608766	GOSR2	0.86
rs1327235	20	10867030	10967030	rs1327235	JAG1	0.46
rs6015450	20	57134512	57234512	rs6015450	GNAS,EDN3	0.12
MTHFR	1	11763367	11794564	rs17367504	MTHFR,NPPB	0.15
ULK4	3	41258094	41983926	rs3774372	ULK4	0.83
rs1458038	4	81333747	81433747	rs1458038	FGF5	0.29
CACNB2	10	18464612	18875804	rs1813353	CACNB2	0.68
C10orf107	10	63087725	63201530	rs4590817	C10orf107	0.84
NT5C2	10	104830930	104948046	rs11191548	CYP17A1,NT5C2	0.91
PLEKHA7	11	16751418	16997566	rs381815	PLEKHA7	0.26
ATP2B1	12	88500959	88632208	rs17249754	ATP2B1	0.84
SH2B3	12	110323135	110378810	rs3184504	SH2B3	0.47
rs10850411	12	113822179	113922179	rs10850411	TBX5,TBX3	0.7
rs1378942	15	72814420	72914420	rs1378942	CYP1A1,ULK3	0.35
ZNF652	17	44716567	44799834	rs12940887	ZNF652	0.38

3.4 Marginal association analysis of the 29 regions in MESA

Table C.7: Marginal association between sBP and the 29 regions in MESA. Four ethnic groups were combined using Fisher's method. MinP: minimum p-value test using GEE. SKAT: the SKAT test using the average value of the outcome. LGRF-G/J: the longitudinal genetic random field model proposed by He et al. (2015). The results are from He et al. (2015). Age, gender, BMI and top 2 PCs were adjusted as covariates.

	name	chr	start	end	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	0.12662	0.11615	0.08422	0.20122
2	MTHFR	1	11763367	11794564	0.44532	0.27329	0.25608	0.19276
3	rs13082711	3	27462913	27562913	0.04979	0.00129	0.00087	0.00359
4	MECOM	3	170278981	170869100	0.84271	0.16447	0.26964	0.25385
5	ULK4	3	41258094	41983926	0.93729	0.94506	0.96538	0.90941
6	SLC39A8	4	103386221	103576438	1.00000	0.95245	0.88354	0.98117
7	GUCY1A3	4	156802313	156877951	0.27964	0.14522	0.14894	0.16461
8	rs1458038	4	81333747	81433747	0.71395	0.26335	0.20692	0.16141
9	rs1173771	5	32800785	32900785	0.70582	0.38030	0.22572	0.46261
10	rs11953630	5	157727980	157827980	0.99994	0.69307	0.69145	0.83252
11	HFE	6	26190488	26211550	0.58656	0.15250	0.14972	0.22746
12	rs805303	6	31674345	31774345	0.06037	0.28225	0.20031	0.19631
13	rs4373814	10	18409978	18509978	0.61972	0.15804	0.12702	0.27193
14	PLCE1	10	95738736	96083139	0.20492	0.31685	0.38811	0.24973
15	CACNB2	10	18464612	18875804	0.51404	0.02932	0.02661	0.02366
16	C10orf107	10	63087725	63201530	0.13607	0.12543	0.07493	0.04878
17	NT5C2	10	104830930	104948046	0.11075	0.02387	0.03046	0.09402
18	rs7129220	11	10257114	10357114	0.51322	0.58221	0.62433	0.64354
19	ARHGAP42	11	100058594	100371866	0.36079	0.55124	0.35965	0.34083
20	PLEKHA7	11	16751418	16997566	0.30172	0.10267	0.07077	0.04177
21	ATP2B1	12	88500959	88632208	0.99872	0.38877	0.48796	0.48406
22	SH2B3	12	110323135	110378810	0.69064	0.46967	0.36973	0.62359
23	rs10850411	12	113822179	113922179	0.99999	0.96044	0.92497	0.91949
24	FES	15	89222929	89245010	0.96616	0.26903	0.35591	0.55000
25	rs1378942	15	72814420	72914420	0.28300	0.02224	0.02575	0.02478
26	GOSR2	17	42350482	42465002	0.78478	0.49338	0.45838	0.57622
27	ZNF652	17	44716567	44799834	0.98444	0.63856	0.76786	0.88685
28	rs1327235	20	10867030	10967030	0.92765	0.48760	0.48532	0.55831
29	rs6015450	20	57134512	57234512	0.40123	0.24997	0.25722	0.17082

Table C.8: Marginal association between dBP and the 29 regions in MESA. Four ethnic groups were combined using Fisher's method. MinP: minimum p-value test using GEE. SKAT: the SKAT test using the average value of the outcome. LGRF-G/J: the longitudinal genetic random field model proposed by He et al. (2015). The results are from He et al. (2015). Age, gender, BMI and top 2 PCs were adjusted as covariates.

	name	chr	start	end	MinP	SKAT	LGRF-G	LGRF-J
1	MOV10	1	113012286	113049891	0.65418	0.61873	0.56224	0.72836
2	MTHFR	1	11763367	11794564	0.13322	0.00934	0.00557	0.00854
3	rs13082711	3	27462913	27562913	0.00434	0.00041	0.00063	0.00241
4	MECOM	3	170278981	170869100	0.40581	0.04256	0.10761	0.07974
5	ULK4	3	41258094	41983926	0.12987	0.55612	0.56460	0.87757
6	SLC39A8	4	103386221	103576438	0.97592	0.87205	0.69878	0.34243
7	GUCY1A3	4	156802313	156877951	0.82750	0.19607	0.24412	0.47640
8	rs1458038	4	81333747	81433747	0.99279	0.32604	0.26011	0.26061
9	rs1173771	5	32800785	32900785	0.50600	0.82381	0.76341	0.87893
10	rs11953630	5	157727980	157827980	0.90491	0.48791	0.58455	0.46902
11	HFE	6	26190488	26211550	0.78654	0.18630	0.31383	0.39228
12	rs805303	6	31674345	31774345	0.11321	0.04401	0.04039	0.05240
13	rs4373814	10	18409978	18509978	0.86003	0.62699	0.50342	0.67279
14	PLCE1	10	95738736	96083139	1.00000	0.90959	0.90494	0.82606
15	CACNB2	10	18464612	18875804	0.23328	0.07493	0.05311	0.04745
16	C10orf107	10	63087725	63201530	0.02302	0.00146	0.00097	0.00086
17	NT5C2	10	104830930	104948046	1.00000	0.77595	0.68184	0.76498
18	rs7129220	11	10257114	10357114	0.47707	0.08902	0.09892	0.18138
19	ARHGAP42	11	100058594	100371866	0.63454	0.44668	0.29560	0.56021
20	PLEKHA7	11	16751418	16997566	0.70240	0.18356	0.13426	0.11532
21	ATP2B1	12	88500959	88632208	0.50290	0.25040	0.29796	0.15524
22	SH2B3	12	110323135	110378810	0.26183	0.71943	0.63767	0.55636
23	rs10850411	12	113822179	113922179	0.95580	0.90671	0.88300	0.89773
24	FES	15	89222929	89245010	1.00000	0.56643	0.57727	0.63884
25	rs1378942	15	72814420	72914420	0.62576	0.23689	0.16497	0.16428
26	GOSR2	17	42350482	42465002	0.58833	0.32515	0.41807	0.53758
27	ZNF652	17	44716567	44799834	0.93545	0.65153	0.70855	0.76768
28	rs1327235	20	10867030	10967030	1.00000	0.86045	0.92321	0.97171
29	rs6015450	20	57134512	57234512	0.64413	0.20782	0.23223	0.28420

4.4 Sensitivity analyses

Table C.9: Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data. Each cell shows the p-value. EUR: European Americans; HIS: Hispanics. DFFS: Density of favorable food stores. DRF: Density of recreational facilities. PHFA: Perceived Healthy Food Availability. PW: Perceived walkability. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E. MinP: minimum p-value test based on GEE. iSKAT-avg./base.: cross-sectional iSKAT using the average/baseline value of repeated measurements as the outcome. Bonferroni correction threshold is 0.00043.

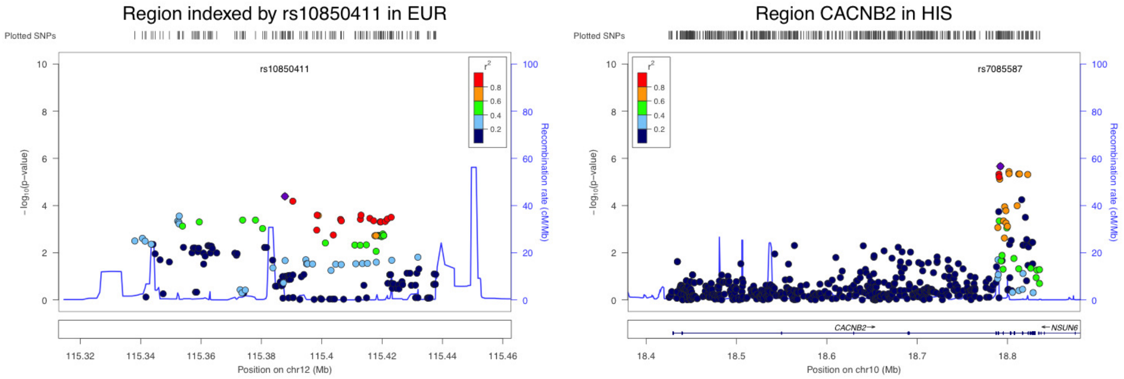
Systolic Blood Pressure - Region Indexed by rs10850411 - EUR								
	Not adjusting for site				Adjusting for site			
	DFFS	DRF	PHFA	PW	DFFS	DRF	PHFA	PW
GE-linear	0.0427	0.7857	0.0005	0.2812	0.0621	0.8518	0.0009	0.2590
GE-spline	0.0570	0.8480	0.0009	0.2127	0.0725	0.8830	0.0010	0.2240
MinP	0.0602	1.0000	0.0047	1.0000	0.0906	1.0000	0.0091	1.0000
iSKAT-avg.	0.2416	0.5134	0.8205	0.9028	0.3352	0.6967	0.8667	0.9001
iSKAT-base.	0.3953	0.7215	0.4331	0.8422	0.5572	0.8872	0.5938	0.8284

Table C.10: Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data. Each cell shows the p-value. EUR: European Americans; HIS: Hispanics. DFFS: Density of favorable food stores. DRF: Density of recreational facilities. PHFA: Perceived Healthy Food Availability. PW: Perceived walkability. GE-base: the proposed test using the baseline measurement of both Y and E. GE-cumulative: the proposed test using the cumulative average of previous measurements. GE: the proposed test as noted in the main text, which assumes full covariate conditional mean (Pepe and Anderson, 1994). $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E are used. Bonferroni correction threshold is 0.00043.

Systolic Blood Pressure								
	Region Indexed by rs10850411 - EUR				CACNB2 - HIS			
	DFFS	DRF	PHFA	PW	DFFS	DRF	PHFA	PW
GE-base	0.4849	0.8250	0.4252	0.7919	0.0783	0.0190	0.9741	0.7015
GE-cumulative	0.1466	0.6995	0.1771	0.5955	0.1094	0.0279	0.3746	0.4476
GE	0.0570	0.8480	0.0009	0.2127	0.1008	0.0346	0.3081	0.3792

3.5 Locus-zoom plots

Figure C.2: Locus-zoom plots of single SNP analysis of systolic blood pressure. The left panel shows the interaction between perceived healthy food availability and region indexed by rs10850411 in European Americans. The right panel shows the interaction between density of recreational facilities and region CACNB2 in Hispanic Americans. Each dot presents the p-value with respect to one SNP in the region.



3.6 Additional data analysis of region CACNB2 and region indexed by rs10850411

Table C.11: Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data: interactions between neighborhood variables and the region CACNB2 on systolic blood pressure. Each cell shows the p-value. EUR: European Americans; AFA: African Americans; HIS: Hispanics; CHN: Asians of Chinese descent. Meta: Meta-analysis combining the results of four ethnic groups using Fisher's combined probability test. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E. MinP: minimum p-value test based on GEE. The working correlation assumed in LGRF is compound symmetric. iSKAT-avg./base.: cross-sectional iSKAT using the average/baseline value of repeated measurements as the outcome. Bonferroni correction threshold is 0.00043.

Systolic Blood Pressure - Region CACNB2 (687 -741 SNPs)										
	Density of favorable food stores					Density of recreational facilities				
	EUR	CHN	AFA	HIS	Meta	EUR	CHN	AFA	HIS	Meta
GE-linear	0.7215	0.1232	0.3837	0.0955	0.1773	0.5592	0.5777	0.7337	0.0753	0.4285
GE-spline	0.7410	0.2133	0.4402	0.1008	0.2708	0.5736	0.6176	0.6703	0.0346	0.2938
MinP	1.0000	1.0000	1.0000	0.0457	0.6282	0.4473	0.2836	1.0000	0.0012	0.0247
iSKAT.avg	0.9833	0.2402	0.3045	0.1023	0.2775	0.7105	0.2057	0.3039	0.0596	0.1572
iSKAT.base	0.9413	0.3396	0.5249	0.1125	0.4394	0.4008	0.2528	0.5071	0.0487	0.1521
	Perceived Healthy Food Availability					Perceived walkability				
	EUR	CHN	AFA	HIS	Meta	EUR	CHN	AFA	HIS	Meta
GE-linear	0.8697	0.5274	0.9331	0.3374	0.8687	0.5087	0.8755	0.5675	0.5535	0.8630
GE-spline	0.8166	0.3934	0.9152	0.3081	0.7784	0.5610	0.9190	0.5241	0.3792	0.8038
MinP	0.4264	1.0000	1.0000	0.8478	0.9799	1.0000	1.0000	1.0000	1.0000	1.0000
iSKAT.avg	0.8886	0.0381	0.2958	0.0278	0.0373	0.3447	0.3270	0.3881	0.1459	0.2575
iSKAT.base	0.9905	0.0174	0.3402	0.6535	0.1945	0.3580	0.5804	0.3571	0.2561	0.4407

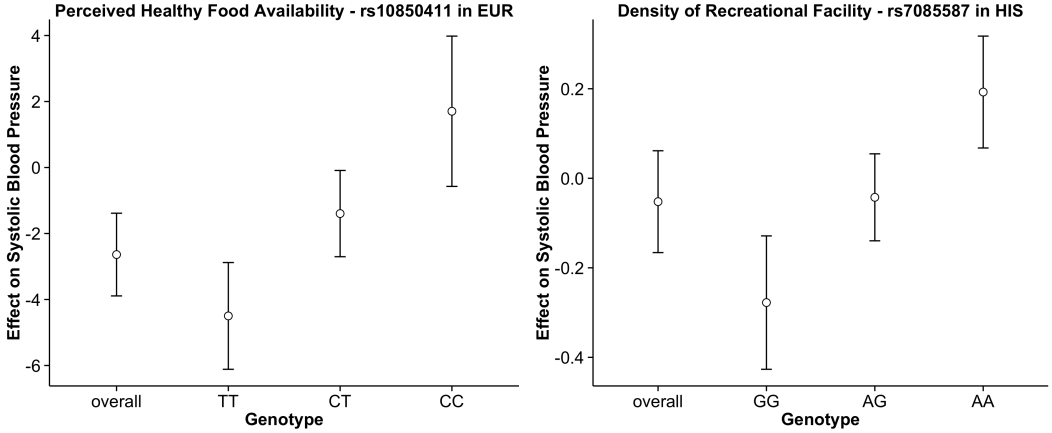
Table C.12: Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data. Each cell shows the p-value. DFFS: Density of favorable food stores. DRF: Density of recreational facilities. PHFA: Perceived Healthy Food Availability. PW: Perceived walkability. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using $\sqrt{n}$ natural cubic-spline basis functions for the main effect of E. MinP: minimum p-value test based on GEE. iSKAT-avg.: cross-sectional iSKAT using the average value of repeated measurements as the outcome. iSKAT-base.: cross-sectional iSKAT using the baseline value of repeated measurements as the outcome. Bonferroni correction threshold is 0.00043.

Region Indexed by rs10850411 - European Americans								
	Systolic blood pressure				Diastolic blood pressure			
	DFFS	DRF	PHFA	PW	DFFS	DRF	PHFA	PW
GE-linear	0.0427	0.7857	0.0005	0.2812	0.3865	0.7021	0.0844	0.9470
GE-spline	0.0570	0.8480	0.0009	0.2127	0.3644	0.7238	0.0527	0.9179
MinP	0.0602	1.0000	0.0047	1.0000	1.0000	1.0000	0.8297	1.0000
iSKAT-avg .	0.2416	0.5134	0.8205	0.9028	0.9147	0.8799	0.6560	0.5510
iSKAT-base.	0.3953	0.7215	0.4331	0.8422	0.7211	0.9039	0.9851	0.7764

3.7 Single SNP GEI analysis of the identified SNPs

The following figure shows the subgroup effect of the neighborhood exposures on systolic blood pressure, stratified by the top SNPs in the region indexed by $rs10850411$ and region $CACNB2$. The overall effects show that higher healthy food availability is associated with lower systolic blood pressure in EUR with a coefficient -2.64 and 95% CI (-3.89, -1.39); Higher density of recreational facility is associated with lower systolic blood pressure in HIS with a coefficient -0.05 and 95% CI (-0.17, 0.06). The effects are modified by $rs10850411$ and $rs7085587$. For example, higher healthy food availability has a larger negative association with systolic blood pressure in European Americans with genotype TT on $rs10850411$ (coefficient = -4.50, 95% CI = (-2.88, -6.12)), but the association is not significant for those with genotype CC (coefficient = 1.70, 95% CI = (-0.57, 3.98)); Higher density of recreational facility has a larger negative association with systolic blood pressure in Hispanic Americans with genotype GG on $rs7085587$ (coefficient = -0.28, 95% CI = (-0.13, -0.43)), but the association is positive for those with genotype AA (coefficient = 0.19, 95% CI = (0.07,0.32)).

Figure C.3: The left panel shows the subgroup effect of perceived healthy food availability on systolic blood pressure stratified by the genotype of rs10850411 in European Americans. The right panel shows the effect of density of recreational facilities on systolic blood pressure stratified by the genotype of rs7085587 in Hispanic Americans.



3.8 Genome-wide GEI analysis of MESA

Figure C.4: Genome-wide set-based analysis of systolic blood pressure in MESA using the proposed method. Four ethnic groups' p-values were combined using Fisher's method for a meta-analysis after an ethnicity specific analysis. Each dot presents the p-value of one gene.

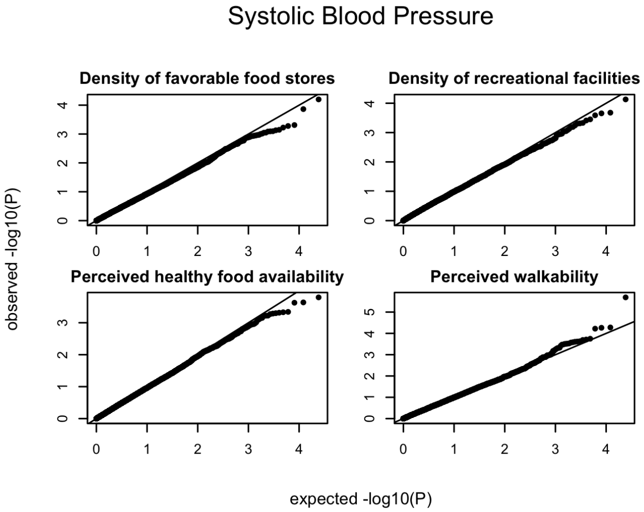


Figure C.5: Genome-wide set-based analysis of diastolic blood pressure in MESA using the proposed method. Four ethnic groups' p-values were combined using Fisher's method for a meta-analysis after an ethnicity specific analysis. Each dot presents the p-value of one gene.

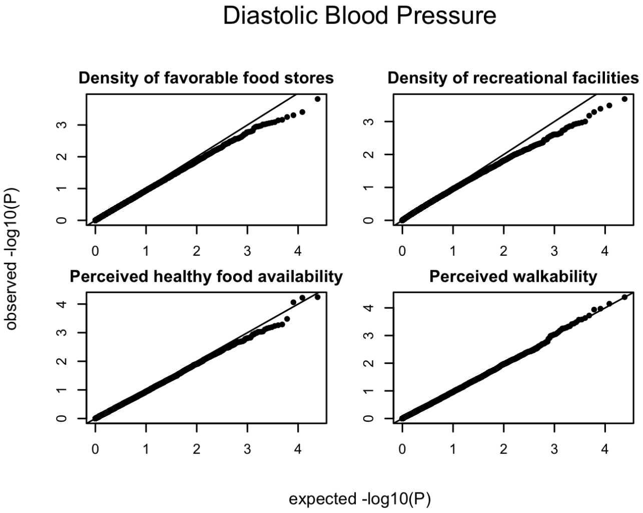


Figure C.6: Genome-wide set-based analysis of systolic blood pressure in MESA using iSKAT with the baseline value of repeated measures. Four ethnic groups' p-values were combined using Fisher's method for a meta-analysis after an ethnicity specific analysis. Each dot presents the p-value of one gene.

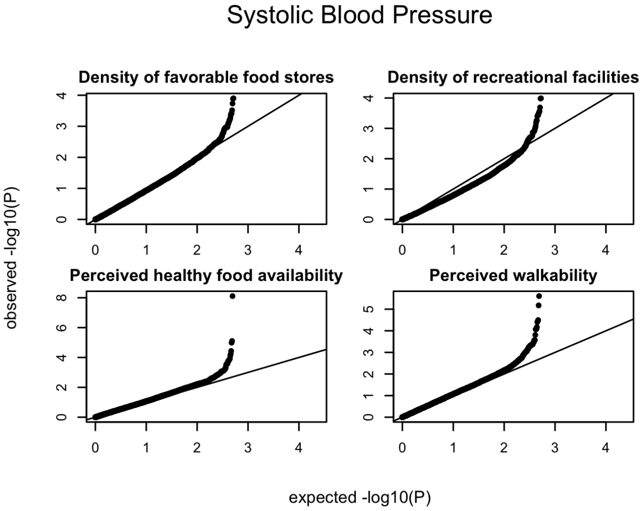
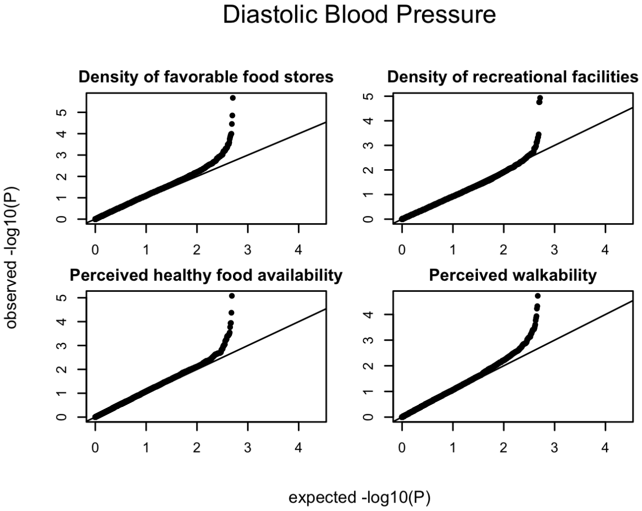


Figure C.7: Genome-wide set-based analysis of diastolic blood pressure in MESA using iSKAT with the baseline value of repeated measures. Four ethnic groups' p-values were combined using Fisher's method for a meta-analysis after an ethnicity specific analysis. Each dot presents the p-value of one gene.



BIBLIOGRAPHY

BIBLIOGRAPHY

Adler, R. J. and Taylor, J. E. (2009), *Random fields and geometry*, Springer Science & Business Media.

Akaike, H. (1998), “Information theory and an extension of the maximum likelihood principle,” in *Selected Papers of Hirotugu Akaike*, Springer, pp. 199–213.

Ballard, D. H., Cho, J., and Zhao, H. (2010), “Comparisons of multi-marker association methods to detect association between a candidate region and disease,” *Genetic epidemiology*, 34, 201–212.

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

Besag, J. (1974), “Spatial interaction and the statistical analysis of lattice systems,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 192–236.

— (1975), “Statistical analysis of non-lattice data,” *The statistician*, 179–195.

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Roux, A. V. D., Folsom, A. R., Greenland, P., JacobsJr, D. R., Kronmal, R., Liu, K., et al. (2002), “Multi-ethnic study of atherosclerosis: objectives and design,” *American Journal of Epidemiology*, 156, 871–881.

Boonstra, P. S., Mukherjee, B., Gruber, S. B., Ahn, J., Schmit, S. L., and Chatterjee, N. (2016), “Tests for Gene-Environment Interactions and Joint Effects With Exposure Misclassification,” *American journal of epidemiology*, kwv198.

Boos, D. D. (1992), “On generalized score tests,” *The American Statistician*, 46, 327–333.

Boulesteix, A.-L. and Strimmer, K. (2007), “Partial least squares: a versatile tool for the analysis of high-dimensional genomic data,” *Briefings in Bioinformatics*, 8, 32–44.

Browning, J. D., Szczepaniak, L. S., Dobbins, R., Horton, J. D., Cohen, J. C., Grundy, S. M., and Hobbs, H. H. (2004), “Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity,” *Hepatology*, 40, 1387–1395.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006), “Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions,” *The American Journal of Human Genetics*, 79, 1002–1016.

Chen, H., Meigs, J. B., and Dupuis, J. (2014), “Incorporating gene-environment interaction in testing for association with rare genetic variants,” *Human heredity*, 78, 81–90.

Christine, P. J., Auchincloss, A. H., Bertoni, A. G., Carnethon, M. R., Sánchez, B. N., Moore, K., Adar, S. D., Horwich, T. B., Watson, K. E., and Roux, A. V. D. (2015), "Longitudinal Associations Between Neighborhood Physical and Social Environments and Incident Type 2 Diabetes Mellitus: The Multi-Ethnic Study of Atherosclerosis (MESA)," *JAMA Internal Medicine*, 175, 1311–1320.

Cornelis, M. C., Tchetgen, E. J. T., Liang, L., Qi, L., Chatterjee, N., Hu, F. B., and Kraft, P. (2012), "Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes," *American Journal of Epidemiology*, 175, 191–202.

Cressie, N. (2015), *Statistics for spatial data*, John Wiley & Sons.

Cui, J. S., Hopper, J. L., and Harrap, S. B. (2003), "Antihypertensive treatments obscure familial contributions to blood pressure variation," *Hypertension*, 41, 207–210.

Davies, R. B. (1980), "Algorithm AS 155: The distribution of a linear combination of χ 2 random variables," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29, 323–333.

Derkach, A., Lawless, J. F., Sun, L., et al. (2014), "Pooled association tests for rare genetic variants: a review and some new results," *Statistical Science*, 29, 302–321.

Fan, R., Zhang, Y., Albert, P. S., Liu, A., Wang, Y., and Xiong, M. (2012), "Longitudinal association analysis of quantitative traits," *Genetic epidemiology*, 36, 856–869.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Genesis Publishing Pvt Ltd.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012), *Applied longitudinal analysis*, vol. 998, John Wiley & Sons.

Furlotte, N. A., Eskin, E., and Eyheramendy, S. (2012), "Genome-Wide Association Mapping With Longitudinal Data," *Genetic epidemiology*, 36, 463–471.

Gauderman, W. J., Murcray, C., Gilliland, F., and Conti, D. V. (2007), "Testing association between disease and multiple SNPs in a candidate gene," *Genetic epidemiology*, 31, 383–395.

Goeman, J. J., Van De Geer, S. A., and Van Houwelingen, H. C. (2006), "Testing against a high dimensional alternative," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 477–493.

Gotze, F. and Tikhomirov, A. N. (1999), "Asymptotic distribution of quadratic forms," *Annals of probability*, 1072–1098.

Grenander, U. (1981), *Abstract inference*, Wiley New York.

He, Z., Zhang, M., Lee, S., Smith, J. A., Guo, X., Palmas, W., Kardia, S. L., Roux, A. V. D., and Mukherjee, B. (2015), "Set-based tests for genetic association in longitudinal studies," *Biometrics*, 71, 606–615.

He, Z., Zhang, M., Zhan, X., and Lu, Q. (2014), "Modeling and testing for joint association using a genetic random field model," *Biometrics*, 70, 471–479.

ICBP (2011), "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk," *Nature*, 478, 103–109.

Kaiser, P., Mujahid, M., Carnethon, M., Bertoni, A., Adar, S., Shea, S., McClelland, R., Lisbeth, L., and Diez Roux, A. (2015), "Neighborhood environments and incident hypertension in the Multi-Ethnic Study of Atherosclerosis," *American Journal of Epidemiology, forthcoming*.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., et al. (2003), "The UCSC genome browser database," *Nucleic Acids Research*, 31, 51–54.

Laird, N. M. and Lange, C. (2010), *The fundamentals of modern statistical genetics*, Springer Science & Business Media.

Lee, S., Wright, F. A., and Zou, F. (2011), "Control of population stratification by correlation-Selected principal components," *Biometrics*, 67, 967–974.

Li, M., He, Z., Zhang, M., Zhan, X., Wei, C., Elston, R. C., and Lu, Q. (2014), "A generalized genetic random field method for the genetic association analysis of sequencing data," *Genetic Epidemiology*, 38, 242–253.

Li, S., Cui, Y., et al. (2012), "Gene-centric gene–gene interaction: A model-based kernel machine method," *The Annals of Applied Statistics*, 6, 1134–1161.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Lin, X., Lee, S., Christiani, D. C., and Lin, X. (2013), "Test for interactions between a genetic marker set and environment in generalized linear models," *Biostatistics*, 14, 667–681.

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., and Lin, X. (2016), "Test for rare variants by environment interactions in sequencing association studies," *Biometrics*, 72, 156–164.

Mallows, C. L. (1973), "Some comments on Cp," *Technometrics*, 15, 661–675.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009), "Finding the missing heritability of complex diseases," *Nature*, 461, 747–753.

Marceau, R., Lu, W., Holloway, S., Sale, M. M., Worrall, B. B., Williams, S. R., Hsu, F.-C., and Tzeng, J.-Y. (2015), "A Fast Multiple-Kernel Method With Applications to Detect Gene-Environment Interaction," *Genetic Epidemiology*, 39, 456–468.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007), "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nature Genetics*, 39, 906–913.

McKusick, V. A. (1998), *Mendelian inheritance in man: a catalog of human genes and genetic disorders*, vol. 1, JHU Press.

Moore, L. V., Roux, A. V. D., Nettleton, J. A., and Jacobs, D. R. (2008), "Associations of the Local Food Environment with Diet Quality—A Comparison of Assessments based on Surveys and Geographic Information Systems The Multi-Ethnic Study of Atherosclerosis," *American Journal of Epidemiology*, 167, 917–924.

Morris, A. P. and Zeggini, E. (2010), "An evaluation of statistical approaches to rare variant analysis in genetic association studies," *Genetic epidemiology*, 34, 188–193.

Mujahid, M. S., Roux, A. V. D., Morenoff, J. D., Raghunathan, T. E., Cooper, R. S., Ni, H., and Shea, S. (2008), "Neighborhood characteristics and hypertension," *Epidemiology*, 19, 590–598.

Mukherjee, B., Ahn, J., Gruber, S. B., and Chatterjee, N. (2012), "Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons," *American Journal of Epidemiology*, 175, 177–190.

Newey, W. K. (1997), "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

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

Papas, M. A., Alberg, A. J., Ewing, R., Helzlsouer, K. J., Gary, T. L., and Klassen, A. C. (2007), "The built environment and obesity," *Epidemiologic Reviews*, 29, 129–143.

Pfeufer, A., van Noord, C., Marciante, K. D., Arking, D. E., Larson, M. G., Smith, A. V., Tarasov, K. V., Müller, M., Sotoodehnia, N., Sinner, M. F., et al. (2010), "Genome-wide association study of PR interval," *Nature Genetics*, 42, 153–159.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L.-J., and Sunyaev, S. R. (2010), "Pooled association tests for rare variants in exon-resequencing studies," *The American Journal of Human Genetics*, 86, 832–838.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., and Willer, C. J. (2010), "LocusZoom: regional visualization of genome-wide association scan results," *Bioinformatics*, 26, 2336–2337.

Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving generalised estimating equations using quadratic inference functions," *Biometrika*, 87, 823–836.

Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H., and Cohen, J. C. (2007), "Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL," *Nature genetics*, 39, 513–516.

Rotar, V. (1974), "Some limit theorems for polynomials of second degree," *Theory of Probability & Its*

Applications, 18, 499–507.

Sallis, J. F., Floyd, M. F., Rodríguez, D. A., and Saelens, B. E. (2012), “Role of built environments in physical activity, obesity, and cardiovascular disease,” *Circulation*, 125, 729–737.

Schwarz, G. (1978), “Estimating the dimension of a model,” *The Annals of Statistics*, 6, 461–464.

Sotoodehnia, N., Isaacs, A., de Bakker, P. I., Dörr, M., Newton-Cheh, C., Nolte, I. M., Van der Harst, P., Müller, M., Eijgelsheim, M., Alonso, A., et al. (2010), “Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction,” *Nature Genetics*, 42, 1068–1076.

Stefanski, L. A. and Boos, D. D. (2002), “The calculus of M-estimation,” *The American Statistician*, 56, 29–38.

Tchetgen Tchetgen, E. J. and Kraft, P. (2011), “On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified,” *Epidemiology*, 22, 257–261.

Thomas, D. (2010), “Gene–environment-wide association studies: emerging approaches,” *Nature Reviews Genetics*, 11, 259–272.

Tzeng, J.-Y., Zhang, D., Chang, S.-M., Thomas, D. C., and Davidian, M. (2009), “Gene-Trait Similarity Regression for Multimarker-Based Association Analysis,” *Biometrics*, 65, 822–832.

Tzeng, J.-Y., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M. I., Sale, M. M., Worrall, B. B., Hsu, F.-C., Thomas, D. C., and Sullivan, P. F. (2011), “Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression,” *The American Journal of Human Genetics*, 89, 277–288.

VanderWeele, T. J., Ko, Y.-A., and Mukherjee, B. (2013), “Environmental confounding in gene-environment interaction studies,” *American Journal of Epidemiology*, 178, 144–152.

VanderWeele, T. J. and Shpitser, I. (2013), “On the definition of a confounder,” *The Annals of Statistics*, 41, 196–220.

Vansteelandt, S., VanderWeele, T. J., Tchetgen, E. J., and Robins, J. M. (2008), “Multiply robust inference for statistical interactions,” *Journal of the American Statistical Association*, 103, 1693–1704.

Voorman, A., Lumley, T., McKnight, B., and Rice, K. (2011), “Behavior of QQ-plots and genomic control in studies of gene-environment interaction,” *PloS One*, 6, e19416.

Wang, L. (2011), “GEE analysis of clustered binary data with diverging number of covariates,” *The Annals of Statistics*, 39, 389–417.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010), “Powerful SNP-set analysis for case-control genome-wide association studies,” *The American Journal of Human Genetics*, 86, 929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011), “Rare-variant association testing for sequencing data with the sequence kernel association test,” *The American Journal of Human Genetics*, 89, 82–93.

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