BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

GENETIC ASSOCIATION METHODS FOR MULTIPLE TYPES OF TRAITS IN FAMILY SAMPLES

by

SHUAI WANG

B.S., Fudan University, 2009
M.S., University of Virginia, 2011

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2015

Approved by

First Reader _____

Josée Dupuis, Ph.D.
Professor of Biostatistics

Second Reader _____

Paola Sebastiani, Ph.D.
Professor of Biostatistics

Third Reader _____

Jame B. Meigs, MD.
Professor of Medicine, Harvard Medical School

DEDICATION

I would like to dedicate this work to my beloved late grandmother Yunhua Xu.

ACKNOWLEDGMENTS

I would like to thank everyone in the past 20 years for your encouragement and enlightenment, without which I wouldn't have been able to accomplish this long educational journey.

I would like to firstly thank my mentor Josée Dupuis, who is one of the most diligent, brilliant and responsible person I've ever met. She's always patient to help with any questions, like a mother. I wouldn't have made so much good progress in the research life without her knowledgeable expertise and guidance. I am grateful to Dr. Dan Keenan at the University of Virginia, without whose vital inspiration, encouragement and enlightenment I wouldn't have been motivated to pursue a Ph.D. degree. My special thanks goes to Dr. John Carulli in Biogen Idec, who's the supervisor during my internship in Biogen Idec where I received useful training in industry. Thank you so much for your generous support and expertise!

For sure my research life wouldn't have been so smooth and so efficient without the company of my classmates and the guidance of my dissertation committee members. I would like to thank Dr. Paola Sebastioni, Dr. James Meigs, Dr. Kathryn Lunetta, Dr. Ching-Ti Liu for their insightful advice on research, teaching and future career. I also appreciate my academic advisor Dr. Serkalem Demissee for designing a flexible course plan, so that I can make the most out of a limited period of time. I also want to thank my

friends: Jae, Wei, Dennis, Jacqui, Yuning and Chen. I feel really fortunate to be friends with you and I wish all of you a brighter future and all your dreams come true.

My deepest gratitude goes to my parents for their constant unconditional love, without which I wouldn't have been able to keep running so persistently. They are always by my sides, believing in me, encouraging me and respecting all my choices.

GENETIC ASSOCIATION METHODS FOR MULTIPLE TYPES OF TRAITS IN FAMILY SAMPLES

SHUAI WANG

Boston University Graduate School of Arts and Sciences, 2015

Major Professor: Josée Dupuis, Professor of Biostatistics

ABSTRACT

Statistical association tests of quantitative traits have been widely used in the past decade, to locate loci associated with a disease trait. For instance, Genome Wide Association Studies (GWAS) have led to tremendous success in finding susceptible genes or associated loci. However, most of the past studies were based on unrelated samples focusing on quantitative or qualitative traits. The analysis of polychotomous traits in family samples is very challenging. This dissertation describes three projects related to methods to conduct association tests beyond continuous traits, such as multinomial traits, bivariate traits, and tests involving haplotypes. The first project focuses on developing a statistical approach to test the association between common or low-frequency variants with a multinomial trait in family samples. It is an important issue because there is no computer efficient software available for this type of question. We employ Laplace approximation in conjunction with an efficient grid-search strategy to obtain an approximate maximum log-likelihood function and the Maximum Likelihood Estimate (MLE) of the variance component. We also successfully incorporate the kinship matrix to adjust for the familial correlation, based on a regression framework. Extensive simulation studies are performed to evaluate the type-I error rate and power in scenarios with causal

variant with different Minor Allele Frequency (MAF). In the second project, we propose an approach to test the association between a genetic variant and a bivariate trait arising from a combination of a quantitative and a binary trait in family samples, based on Extended Generalized Estimating Equations (EGEE). Multiple phenotype-genotype association tests are often reduced to univariate tests, decreasing efficiency and power. Our approach is shown to be much more powerful and efficient than univariate association tests adjusted for multiple testing. The third project involves the development of a general framework for meta-analysis of haplotype association tests, applicable to both unrelated and family samples. Although meta-analysis has been widely used in single-variant and gene-based tests, there are few existing methods to meta-analyze haplotype association tests. A predominant advantage of our novel approach is that it accommodates cohort-specific haplotypes as well as haplotypes common to all cohorts. The cohort participants may be either related or unrelated. Our approach consists of two stages: in the first stage, each cohort performs a haplotype association test, reports the estimates of effect size, variance, haplotypes, and their frequency. In the second stage, a generalized least square method is applied to combine the results of all the cohorts into one vector of meta-analysis coefficients. Our approach is shown to have the correct type-I error rate in scenarios with different between and within cohort variation. We also present an application to exome-chip data from a large consortium. Through the three projects, we are able to tackle the problem of conducting association tests for non-continuous traits in family samples. All the approaches achieve the correct type-I error rate and are computationally efficient. We hope these approaches will not only facilitate analyses of

categorical traits in family samples, but will also provide a basis for future methodological development of statistical approaches for non-continuous traits.

Contents

List of Figures

List of Tables

List of Abbreviations

AGQ	Adaptive Gaussian Quadrature
BMI	Body Mass Index
BMR	Basal Metabolic Rate
CHARGE	Cohorts of Heart and Aging Research Genomic Epidemiology
DGI	Diabetes Genetics Initiative
DIAGRAM	Diabetes Genetics Replication and Meta-analysis
EGEE	Extended Generalized Estimating Equations
FG	Fasting Glucose
FHS	Framingham Heart Study
FI	Fasting Insulin
FUSION	Finland-United States Investigation of NIDDM Genetics
GDT	Generalized Disequilibrium Test
GEE	Generalized Estimating Equations
GLMM	Generalized Linear Mixed Model
GWAS	Genome Wide Association Studies
IBD	Identity-by-Descent
LME	Linear Mixed Effects Model
LRT	Likelihood Ratio Test
MACH	Markov Chain based Haplotyper
MAF	Minor Allele Frequency
MANOVA	Multivariate Analysis of Variance
min P	minimum p-value
MLE	Maximum Likelihood Estimator
PCl	Penalized Conditional log-likelihood
SHARe	SNP Health Association Resource
sibTDT	sib Transmission Disequilibrium Test
SKAT	Sequence Kernel Association Test
TATES	Trait-based Association Test that used Extended Simes procedure
T2D	Type-2 Diabetes
TDT	Transmission Disequilibrium Test
WTCCC	Wellcome Trust Case Control Consortium

Chapter 1

Introduction

1.1 Overview

Statistical association tests for quantitative traits have been widely used in the past decade, to locate loci associated with a disease of interest and to better understand the genetic architecture underlying a disease. Genome Wide Association Studies (GWAS) have led to tremendous successes in finding susceptible genes or associated loci. For instance, Dupuis et al. 2010 [10] identified 9 new loci associated with Fasting Glucose (FG), 1 new loci associated with Fasting Insulin (FI) and Homeostasis Model Assessment of Insulin Resistance. Scott et al. 2012 [39] increased the total number of variants associated with glycemic traits to 53, of which 33 also increased type 2 diabetes risk. However, most of the past studies were based on unrelated samples focusing on quantitative or binary traits. Analysis of polychotomous traits in family samples remains a challenge. Available approaches are limited to study a categorical trait with more than 2 categories, for example, diabetes status with 3 categories (non-diabetic (FG <=5.6mmol/L), impaired fasting glucose tolerant (5.6mmol/L< FG <=6.9mmol/L) and diabetic (FG >7 mmol/L)) in family samples. To address this issue, we develop an approach to test the association between genetic variants and multinomial traits in family samples, and describe our novel

approach in Chapter 2.

Univariate association tests have been the main theme in GWAS or GWAS meta-analysis due to the ease of implementation and superb computer efficiency. Yang et al. 2010 [52] proposed a method to combine unvariate tests to perform a test of association with multivariate phenotypes. This method seems computer efficient and applicable to both family and unrelated samples, quantitative and categorical traits, but it requires prior knowledge of the covariance matrix of the test statistics, which is not always available or difficult to estimate from external sources. When both traits are quantitative, the joint association test can be easily derived from univariate association tests, in the framework of linear mixed effects model (LME). However, when one trait is categorical, for example, binary, it is not feasible to derive a test statistic in closed form, because the likelihood function doesn't have a closed form in the presence of random effects. Considering all the existing limitations mentioned above, we propose a method to conduct joint association test of bivariate phenotypes in family samples (Chapter 3), with no stringent restrictions nor assumptions. Simulation studies show our methods achieve the correct type-I error rate in the scenarios evaluated, is computationally feasible and is more powerful than existing methods.

The third topic is meta-analysis of haplotype association test in family samples and unrelated samples. In recent years, meta-analysis has been widely used in genetic epidemiology, as a way to maximize sample size and to improve power. Several statistical methods have been proposed to meta-analyze single-variant tests, gene-based tests [24], and gene-environmental interaction tests [27]. Through meta-analysis, some weaker signals which would not have been found in single study were discovered. However, there are no existing methods for meta-analysis of haplotype association tests, most likely

because of a few challenges. Firstly, haplotypes observed by different cohorts or ethnicity groups can vary a lot; secondly, the haplotype structure can be very complex, especially in a region containing a large number of variants. We propose a two-stage approach which overcomes these hurdles and accommodates cohort-specific haplotypes in addition to haplotypes observed in all cohorts. Simulation studies demonstrate that our approach has the correct type-I error rate in the scenarios evaluated, is computationally efficient and can be more powerful than gene-based and single-SNP tests for some underlying genetic models.

1.2 Association tests for multinomial traits

1.2.1 Unrelated samples

GWAS have proved to be very successful in identifying important genetic components underlying a binary disease trait. Prior to 2010, GWAS have discovered 38 SNPs associated with Type-2 Diabetes (T2D) status, in addition to 2 dozen loci associated with glycemic traits [2]. The Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium which was formed by combining association results from the Wellcome Trust Case Control Consortium (WTCCC), the Finland-United States Investigation of NIDDM Genetics (FUSION) group, and the Diabetes Genetics Initiative (DGI), has eventually led to the discovery of many common variants with small effect size associated with T2D status, a binary trait [55]. Among the findings, researchers used to analyze dichotomous T2D status. However, in the clinical diagnosis of T2D, there is a FG gap (5.6mmol/L$<FG<=$6.9mmol/L) called impaired glucose tolerance and samples within this range may be omitted or combined with "non-T2D", reducing the specificty of "normal"category. Making use of the glucose impaired individuals may improve power. For example, a categorical trait could be defined with the following three categories: diabetic, pre-diabetic, and non-diabetic. Compared to the commonly used binary T2D

status, this multinomial categorization better captures the original distribution of FG level and T2D risk, and retains a larger sample size. Another example arises from multiple categorical outcomes. Assume in a study, study participants have a clear diagnosis of diabetes v.s. no diabetes, hypertension v.s. no hypertension. A potential research interest is to study the association between the genetic variants and the joint phenotype by modeling the two outcomes as one categorical variable, e.g. a three-category variable: diabetes & hypertension, diabetes & no hypertension, and no diabetes. Software implementation for such genetic association testing for a multi-category trait is readily available. For example, the R function "glm" can be used to perform the multinomial association tests for unrelated samples:

$log\frac{P(Y=k)}{P(Y=3)} = \alpha_k + \beta_k X + \gamma_k G$

where $k = 1, 2$; Y is a three-category outcome variable; X is a matrix of covariates, and G is the genotype. The null hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$, is often tested by a Likelihood Ratio Test (LRT) of the form of $-2log\frac{L_0}{L_a}$ (L_0 denotes the maximum likelihood function under H_0; L_a denotes the maximum likelihood function under $H_0 \cup H_a$) with an asymptotic χ^2_2. In the context of a three-category outcome, this multinomial logit model has twice the number of parameters to estimate, compared to the logistic model for the binary trait, and therefore requires a larger sample for accurate estimation. When sample size is small and the categories are ordered, ordinal regression could serve as another option, with the following model (using the same notation as in the general logit model):

$log\frac{P(Y<=k)}{1-P(Y<=k)} = \alpha_k + \beta X + \gamma G$ where $k = 1, 2$;

This ordinal model only has one more parameter α_k to estimate compared to the logistic model for a binary outcome.

When neither sample size or computation is a concern, model fit can be used to select which model to use. Deviance is an index commonly used to evaluate model fit and

is defined as: $D = -2log\frac{L_0}{L_{saturated}}$ ($L_{saturated}$ is the maximum likelihood function of the saturated model). Smaller deviance is usually preferred over larger deviance.

1.2.2 Family samples

Testing the association of categorical traits in family samples becomes more complicated. The Transmission Disequilibrium Test (TDT) was one of the first family-based association tests proposed [42]. In the presence of n trios (two parents and 1 affected child), the TDT measures the over transmission of a particular allele from heterozygous parents to the affected offspring.

A few methods and software were developed for binary traits in family samples [26], among which some were based on TDT, such as sibTDT [41]. One of the major restrictions is that it does not allow for inclusion of covariate adjustments. There are some other family-based methods based on GEE framwork. The Generalized Disequilibrium Test (GDT) developed by Chen et al. 2009 [5] was a great advance, in terms of model flexibility and generality. The GDT can accommodate large and general pedigrees, while adjusting for covariates and incorporating family weights. However, it can not be directly applied to multinomial traits.

Genetic association approaches for population-based (unrelated) samples can be adapted to family-based samples by modeling family correlation as a function of the kinship in the framework of a LME.

The kinship matrix measures the pairwise kinship distance in a general pedigree. Given a general pedigree with n subjects, the $n \times n$ kinship matrix can be easily obtained. There are four required columns in a pedigree file: family id, id, father id and mother id. For example (Table 1.1):

Table 1.1: Pedigree Example

famid	id	father	mother
1	1	0	0
1	2	0	0
1	3	1	2
1	4	1	2
2	5	0	0
2	6	0	0
2	7	5	6

where 0 in father and mother fields denotes that the subject is a founder. The definition of kinship coefficient ϕ_{ij} is the probability that an allele randomly selected from individual i and an allele randomly selected from the same locus of individual j are identical-by-descent (IBD). The coefficient of relationship is defined as twice the coefficient of kinship (Table 1.2).

Table 1.2: Example: coefficient of kinship and coefficient of relationship

Relationship	**coefficient of kinship**	**coeffcient of relationship**
Self	0.5	1
Monozygotic twins	0.5	1
Parent-child	0.25	0.5
Full siblings	0.25	0.5
Half siblings	0.125	0.25
First cousins	0.0625	0.125
Unrelated	0	0

We develop an association test for multinomial traits in the framework of generalized LME, allowing for covariate adjustment while accounting for the familial correlation, in the form of a kinship matrix (Chapter 2). Laplace approximation is first applied to the multiple integrals to approximate the marginal likelihood in closed form. Then grid search in combination with Newton-Ralphson algorithm is proposed to calculate the Maximum Likelihood Estimator (MLE) of the variance component efficiently. We perform extensive simulation studies to first evaluate the type-I error rate of both our approach and GLMM and then compare the power of both methods. Given both models have the same correct type-I error rate, it's reasonable to conclude our model is more powerful than GLMM in all the scenarios evaluated. We also present an application to assess the association between FHS SNP Health Association Resource (SHARe) gentoypes and a three-category BMI variable among FHS participants.

1.3 Joint association tests for bivariate phenotypes

Most published GWAS analyses are univariate or reduced to univariate in testing the association with multiple phenotypes. When multiple phenotype-genotype associations are assessed, univariate tests are easy to implement. However, univariate testing suffers a loss in both efficiency and power. Several methodological approaches have been developed for different types of traits. We review the existing methods with respect to strength and limitations starting from quantitative traits, followed by binary traits.

1.3.1 Quantitative traits

One challenge of testing the association for two quantitative traits arises when it is applied to two correlated phenotypes in family samples. A good joint test should adjust for the correlation between the two phenotypes, while taking into consideration the familial correlation at the same time. There are several different methods proposed.

MANOVA (Multivariate Analysis of Variance) is the most basic joint association test, as is an extension of univariate LME. Multiphen proposed by O'Reilly et al. 2012 [34] used ordinal regression to model the genotypes as a function of a collection of phenotypes of any type (quantitative, binary, oridinal) in unrelated samples.

O'Reilly and colleagues assume that $Y_i = (Y_{i1}, \cdots, Y_{iK})$ is a vector containing K phenotypes for the i^{th} individual; X_{ig} is the additively coded genotypes, i.e. $X_{ig} \in \{0, 1, 2\}$. The univariate association model for trait k is simply: $Y_{ik} = \alpha_k + \beta_{gk} X_{ig} + \epsilon_{ik}$

In the Multiphen approach, the typical regression model is inverted by means of an ordinal regression, such that

$log \frac{P(X_{ig} <= m)}{1 - P(X_{ig} <= m)} = \alpha_m + \sum \beta_k Y_{ik}$

where m can be selected as 0 or 1. This approach is limited in two aspects: the same covariates are used for the K phenotypes and it is applicable to unrelated samples only. Sluis et al. 2013 [45] developed a method called Trait-based Association Test that used Extended Simes procedure (TATES) to efficiently analyze multivariate phenotype-genotype association for GWAS. They denote $p_{(1)}, \cdots, p_{(m)}$ as the ascendingly ordered p-value for association of each of the m traits with a genetic variant. In TATES, the m p-values are combined into a single p-value defined as $p_T = min(\frac{m_e p_{(j)}}{m_{ej}})$ where m_e is the estimated number of independent p-values out of the m traits, and m_{ej} is the estimated number of independent p-values out of the first j p-values.

Yang et al. 2010 [52] proposed a method to combine univariate association tests based on a method originally proposed by O'Brien [32]. O'Brien and colleagues denote $T = (T_1, \cdots, T_m)$ to be the test statistics of the m traits, and assume that T follows a multivariate normal distribution with a mean of $\beta = (\beta_1, \cdots, \beta_m)$ and a covariance

matrix of Σ. The global null hypothesis is $H_0 : \beta = 0$. O'Brien et al. proposed to use the test statistic $e^T\Sigma^{-1}T$ which followed a normal distribution with a mean of 0 and a covariance matrix of $e^T\Sigma^{-1}e$. where e is the uniform weight $(1, \cdots, 1)$ imposed on the test statistics. Yang et al. proposed to use non-uniform weights to reflect potential heterogenity and construct the test statistic as $T_w^T\Sigma^{-1}T$ where T_w and T are the testing statistic based on training and testing samples respectively. This method has the strength of computational simplicity. However, because the covariance matrix Σ is not readily available, it is challenging to obtain a good unbiased estimate of the covariance matrix Σ.

Stephens et al. 2013 [43] proposed a unified framework of multiple phenotype association based on bayesian methods, which seemed very appealing due to its generality, but it might not be as competitive in terms of the computational efficiency when applied to large pedigrees.

1.3.2 Bivariate phenotypes consisting of quantitative and binary traits

Although there have been some advances in the association test for multivariate quantitative traits, the methodology for association test between genetic variants and a mixture of non-continuous traits has not been studied systematically. For the type of problems involving correlated quantitative and binary traits, Generalized Estimating Equations (GEE) might serve as a solution, and prove to be robust to misspecified working correlation matrix. The correlation parameters are treated as nuissance parameters in the estimation, and thus are not estimable. However, covariance matrix and correlation parameters are needed for the purpose of hypothesis testing. Hall et al. [16] developed Extended Generalized Estimating Equations (EGEE) based on quasi-likelihood function. It overcomes the limitations that in GEE the correlation parameters are not estimable, while retaining the many good properties of GEE: the parameter estimation of EGEE

is also robust to the misspecification of covariance matrix, and as efficient as GEE. More importantly, unlike GEE, the correlation parameters are estimated along with the regression coefficients. Liu et al. 2009 [25] proposed to use EGEE to conduct a joint association test for a mixture of binary and quantitative traits in a regression framework for unrelated samples only. EGEE was shown to have the correct type-I error rate and was more powerful than univariate tests adjusted for multiple testing.

Here we propose an approach based on EGEE, to perform a joint association test for a mixture of binary and quantitative traits for samples with familial correlation. Simulation studies demonstrate our approach achieves the correct type-I error rate and can be more powerful, compared to univariate tests adjusted for multiple testing in certain scenarios.

1.4 Meta-analysis of genetic association tests

1.4.1 Meta-analysis of single-variant analysis

Hu et al. 2013 [20] summarized an approach based on score statistics for meta-analyzing genetic association results. In the situation with L independent studies, m variants, the combined score statistic can be obtained as $\overline{U} = \sum_{l=1}^{L} U_l$ where U_l is the association score statistic for study l ($l = 1, \cdots, L$). The covariance matrix is $\overline{V} = \sum_{l=1}^{L} V_l$ where V_l is the covariance matrix of the score statistic of study l. $\overline{U}$ is approximately m-variate normally distributed with a mean of 0 and a covariance of $\overline{V}$. If some studies don't have nor observe a mutation in a site, the corresponding entries of U_l and V_l can be set to 0.

1.4.2 Meta-analysis of gene-based tests

For meta-analysis of gene-based tests, with weights vector w based on MAF (Minor Allele Frequency), the meta-analysis burden score statistic can be written as $U_s = w^T\overline{U}$, with a covariance $V_s = w^T\overline{V}w$. The test statistic $\frac{U_s}{\sqrt{V_s}}$ follows a standard normal distribution.

A similar meta-analysis formulation for the SKAT (Sequence Kernel Association Test) statistic was proposed by Hu et al. The SKAT statistic is simply $\overline{U}^T W\overline{U}$ where W is the diagonal weight matrix with diagonal elements equaling to w based on the MAF and beta density functions. The test statistic has a null distribution of $\sum_{j=1}^{j=m}\lambda_j\chi^2_{1j}$, where λ_j are the eigenvalues of $\overline{V}^{1/2}W\overline{V}^{1/2}$ and the χ^2_{1j} are independent χ^2_1 variables.

Hu proposed some strategies when the score functions or variance estimates were not available. Because the three forms of the univariate association tests (Wald, score, LRT) are equivalent asymptotically, given study l, the score function can be approximated by $U_j = z_j w_j$ where z_j is the z-test statistic equivalent, w_j is the approximation to the square root of variance $\sqrt{V_{jj}}$. The overall covariance matrix of U is $var(U) = WRW$ where R is the correlation matrix of the z-test statistics; W is a diagonal matrix with diagonal elements equaling to the squared root of approximated variance of the score function. Hu and colleagues suggested using the correlation of the genotype for R.

There are no existing methods for the meta-analysis of haplotype association tests. However, for the purpose of fine-mapping and as a follow-up to GWAS or ExomeChip/Sequencing, development of an approach to meet this demand is timely. We propose a novel approach which integrates information from cohorts of any types (either

family-based or unrelated), so that all the haplotypes observed by all cohorts can be combined into one vector of summary coefficients. Based on our framework, both global association test and any single haplotype association test can be easily obtained.

1.5 Dissertation Outline

In this dissertation, we develop statistical approaches to conduct genetic association tests for multiple types of traits in family samples, as well as a general meta-analysis approach of haplotype association tests. Each chapter consists of complete methods section, extensive simulation studies and a real data application. The data analysis of Chapter 4 has been published in Nature communications [49]. As a methods paper, the whole chapter of Chapter 4 with a new data analysis has been submit to plos genetics and is currently under review; the manuscripts of Chapter 2 and 3 are in preparation.

In Chapter 2, we propose an efficient statistical approach to test the association for multinomial traits in family samples. We apply Laplace approximation to approximate the closed form of maximum likelihood function, in conjunction with an efficient grid-search scheme to approximately locate the MLE of the variance component. We evaluate our approach by means of simulation studies in different scenarios: common variants versus low-frequency variants; balanced design versus unbalanced design. We show that our approach has the correct type-I error rate in the scenarios evaluated. Then we compare the power of our approach to the GLMM clustered by families with respect to different MAF, and we show that our approach is consistently more powerful.

In Chapter 3, we develop an approach to jointly test the association for a quantitative trait and a binary trait in family samples. Based on a regression framework with random effects accounting for familial correlation, we use quasi-likelihood based EGEE to generate the

score equations for both regression coefficients and correlation coefficients. We solve the score equations by means of Fisher's scoring algorithm. Extensive simulation studies are performed to assess the type-I error rate in a variety of scenarios with MAF ranging from 0.01 to 0.3. We calculate the power of our approach and compare to univariate association test adjusted for multiple testing using Gao's method recommended by Hendricks et al. Simulation studies show our approach achieves the correct type-I error rate and is more powerful than univariate tests. Lastly, we apply our approach to the Framingham Heart Study (FHS) data, selecting BMI and T2D status as the bivariate phenotypes and perform chromosome-wide association study on chromosome 16. The software and manuscripts are under preparation.

In Chapter 4, we develop a general approach to meta-analyze haplotype association tests from different cohorts. Our method consists of two stages. In the first stage, a haplotype association test is performed at the cohort level; in the second stage, a generalized least square method is applied to combine results from all cohorts, into a vector of meta-analysis coefficients. Our method has quite a few advantages and is flexible: it is applicable to cohorts consisting of either unrelated samples or family samples and it does not put restrictions on the observed haplotypes. In other words, cohorts can contribute cohort-specific haplotypes in addition to those observed in all cohorts. We evaluate the type-I error rate of our approach in several different scenarios, with different between and within cohort variation. Results show our approach has the correct type-I error rate in all the scenarios considered. We compare the power of our approaches to univariate testing of SNP effect adjusted for multiple testing, and our approach is at least as powerful, even in scenarios with single SNP rather than haplotype effects. In the section on data analysis, we apply our approach to a Cohorts of Heart and Aging Research Genomic Epidemiology (CHARGE) exome-chip study focusing on a known T2D associated gene *G6PC2*. The

results are consistent with our prior findings, and are more significant than any single-SNP and gene-based tests in this region. The work of Chapter 4 has been submit to plos genetics, and the website for this software are currently under construction on the website of Boston University medical campus (www.bumc.bu.edu).

In Chapter 5, we summarize the three approaches developed as part of this dissertation, and discuss the pros and cons, in addition to the future direction we want to pursue.

Chapter 2

Multinomial association test in family samples

2.1 Introduction

Genetic association test of continuous phenotypes has led to great success in finding susceptible genes or variants related to a disease. Various methods and efficient software have been developed and used widely. For family samples, due to the correlation between relatives and the violation of the independence assumption in the ordinary linear regression, some alternate approaches were proposed. For example, Therneau and colleagues developed a R package (kinship) to apply LME to conduct association tests between a genetic variant and a continuous trait in family samples. Similar extension to account for familial correlation using mixed effects models has been proposed for gene-based association tests [4] [35]. The progress in family samples mostly applies to quantitative traits. However, methods are needed to study categorical trait(s) with more than 2 categories in family samples. For example, Body Mass Index (BMI) has five generally accepted categories ($BMI < 18.5$ underweight; $18.5 \leq BMI < 24.9$ normal weight; $25.0 \leq BMI \leq 29.9$ overweight; $30 \leq BMI \leq 34.9$ class I obesity; $35 \leq BMI \leq 39.9$ class II obesity; $BMI \geq 40$ class III obesity), and approaches for genetic association analysis of multi-category traits are quite limited.

Wang et al. 2006 [48] proposed a proportional odds logistic model which accommodated covariates. However, there are still a few limitations. First, this approach is restricted to nuclear families and can not handle complex family structure. Second, the relevant software has not been made publicly available. Diao et al. 2010 [9] proposed a general framework for linkage and association tests for ordinal traits. Their method utilized Adaptive Gaussian Quadrature (AGQ) to approximate the maximum log-likelihood and a LRT was performed to test the hypothesis of no association between the genetic variant and the ordinal trait of interest. Again this approach has not been widely used due to the lack of computationally efficient software. Another possible option is the SAS GLMM procedure, which can incorporate a kinship matrix. However, due to the large computational burden, the GLMM procedure is not able to accommodate even small pedigrees. In this chapter, we propose a model to perform genetic association tests for multinomial phenotypes accounting for familial correlation.

2.2 Methods

2.2.1 Adaptive Gaussian Quadrature

The basic model for multinomial trait can be formulated as

$$g(P(Y_{ij} = k|G_{ij}, R_{ij})) = \alpha_k + \beta_k G_{ij} + X_{ij}^T \gamma_k + R_{ij} \tag{2.1}$$

where

k=1, ..., (K-1), and K is the total number of categories of the trait under study;

$j = 1, ..., n_i$ denotes the j^{th} subject of the i^{th} family;

G_{ij} denotes the genotype (e.g., it can be additively coded as 0,1 or 2);

X_{ij} denotes the covariates vector;

$R = \begin{pmatrix} R_{11} \\ \vdots \\ R_{1j} \\ \vdots \\ R_{1n_1} \\ \vdots \\ R_{m_1} \\ \vdots \\ R_{m_{n_m}} \\ \vdots \\ R_{N_{n_N}} \end{pmatrix} \sim N(0, \sigma^2\Sigma_{kin})$ is a random effect vector to account for the familial correlation.

There are a variety of options to use for the link function g. For example, g can take the form of the canonical link function general logit, linear function, probit function, etc. Here general logit is used, such that $g(P(Y_{ij} = k|G_{ij}, R_{ij})) = log(\frac{P(Y_{ij}=k|G_{ij},R_{ij})}{P(Y_{ij}=K|G_{ij},R_{ij})}$ $\forall(k = 1, \cdots, (K-1))$.

The log likelihood function does not have a closed form, due to multiple integration of nonlinear exponential functions. However, we do need a closed form function to estimate the parameters as well as to perform hypothesis testing. I first explore AGQ to approximate the log likelihood function.

The essential idea of AGQ is described below. Suppose that $I = \oint e^{-f(R_i,\theta_i)}dR_i$, where R_i is assumed to have a multivariate normal distribution. A coordinate transformation is applied to the integral:

$$R_i = \widehat{R_i} + (f'')^{-1/2}z_i \tag{2.2}$$

where

$\widehat{R_i} = arg_{R_i}minf(R_i, \theta_i)$ assuming $\theta_i = (\sigma^2, \alpha_1, \cdots, \alpha_{K-1}, \beta_1, \cdots, \beta_{K-1}, \gamma_1, \cdots, \gamma_{K-1})$ is given;

$f'' = f''_{R_i R_i}|_{R_i=\widehat{R_i}}$;

and z_i has a standard multivariate normal distribution.

Then the equivalent form of the likelihood function becomes:

$$\begin{aligned} I &= \oint e^{-f+z_i^2/2} e^{-z_i^2/2} |f''|^{-1/2} dz_i \\ &= |f''|^{-1/2} (2\pi)^{n_i/2} \int \dots \int e^{-f+z_i^2/2} \frac{1}{(2\pi)^{n_i/2}} e^{-z_i^2/2} dz_{i1} \dots dz_{i_{ni}} \\ &\approx |f''|^{-1/2} (2\pi)^{n_i/2} \sum_{j_1=1}^{N_{GQ}} \dots \sum_{j_{n_i}=1}^{N_{GQ}} exp\{-f + \frac{||z_{j_1}, \dots, z_{j_{n_i}}||^2}{2}\} \prod_{k=1}^{n_i} w_{j_k} \end{aligned} \quad (2.3)$$

where $z_{j_1}, \dots, z_{j_{n_i}}$ denotes the pre-specified grid points and w_{j_k} are the weights based on the grid points and the standard multivariate normal distribution.

The marginal likelihood is derived as follows:

$$\begin{aligned} &L(\theta) \\ &= \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} \prod_{k=1}^{K-1} \left(\frac{e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}}{1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}}\right)^{I(Y_{ij}=k)} \left(\frac{1}{1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}}\right)^{I(Y_{ij}=K)} \phi(R_i) dR_i. \end{aligned} \quad (2.4)$$

Define f_i as

$$f_i = -log\{\prod_{j=1}^{n_i} \prod_{k=1}^{K-1} \left(\frac{e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}}{1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}}\right)^{I(Y_{ij}=k)} \left(\frac{1}{1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}}\right)^{I(Y_{ij}=K)} \phi(R_i)\} \quad (2.5)$$

such that

$$L(\theta) = \prod_{i=1}^{n} \int e^{-f_i} dR_i. \quad (2.6)$$

After simplifying the terms of f_i, we obtain

$f_i = -\sum_{j=1}^{n_i} log\{\prod_{k=1}^{K-1}(\frac{e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}}{1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}})^{I(Y_{ij}=k)}(\frac{1}{1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}})^{I(Y_{ij}=K)}\}$
$-log\phi(R_i)$

$= -\sum_{j=1}^{n_i}\{\sum_{k=1}^{K-1} I(Y_{ij} = k)[\alpha_k + \beta_k G_{ij} + \gamma_k X_{ij} + R_{ij} - log(1 + \sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}})] - I(Y_{ij} = K)log(1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}})\} - log\phi(R_i)$

$= \sum_{j=1}^{n_i}\{\sum_{k=1}^{K-1} I(Y_{ij} = k)[log(1 + \sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}) - (\alpha_k + \beta_k G_{ij} + \gamma_k X_{ij} + R_{ij})] + I(Y_{ij} = K)log(1 + \sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}})\} - log\phi(R_i)$

$= \sum_{j=1}^{n_i} log(1 + \sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}) - \sum_{j=1}^{n_i}\sum_{k=1}^{K-1} I(Y_{ij} = k)(\alpha_k + \beta_k G_{ij} + \gamma_k X_{ij} + R_{ij}) - log\phi(R_i)$

$= \sum_{j=1}^{n_i} log(1 + \sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}) - \sum_{j=1}^{n_i}\sum_{k=1}^{K-1} I(Y_{ij} = k)(\alpha_k + \beta_k G_{ij} + \gamma_k X_{ij}) - \sum_{j=1}^{n_i} R_{ij}(1 - I(Y_{ij} = K)) - log\phi(R_i)$

$= \sum_{j=1}^{n_i} log(1 + \sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{ij}+\gamma_k X_{ij}+R_{ij}}) - \sum_{j=1}^{n_i}\sum_{k=1}^{K-1} I(Y_{ij} = k)(\alpha_k + \beta_k G_{ij} + \gamma_k X_{ij}) - \sum_{j=1}^{n_i} R_{ij}(1 - I(Y_{ij} = K)) + \frac{n_i}{2}log2\pi + \frac{1}{2}log|\Sigma_{kin}| + n_i log\sigma + \frac{R_i^T\Sigma_{kin}^{-1}R_i}{2\sigma^2}.$

To implement the coordinate transformation, we take both the first and the second derivative of f_i with respect to R_i:

$\forall m, n = 1, \cdots, n_i, m \neq n$

$$\frac{\partial f_i}{\partial R_{im}} = (f_i)'_m = \frac{\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{im}+\gamma_k X_{im}+R_{im}}}{1 + \sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{im}+\gamma_k X_{im}+R_{im}}} - (1 - I(Y_{im} = K)) + \frac{(\Sigma_{kin}^{-1})_{m\cdot}R_i}{\sigma^2} \tag{2.7}$$

($(\Sigma_{kin}^{-1})_{m\cdot}$ is the m-th row of (Σ_{kin}^{-1}))

$$\frac{\partial^2 f_i}{\partial R_{im}^2} = (f_i)''_{mm} = \frac{\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{im}+\gamma_k X_{im}+R_{im}}}{(1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{im}+\gamma_k X_{im}+R_{im}})^2} + \frac{(\Sigma_{kin}^{-1})_{mm}}{\sigma^2} \tag{2.8}$$

$$(f_i)''_{mn} = \frac{(\Sigma_{kin}^{-1})_{mn}}{\sigma^2} \tag{2.9}$$

Therefore, the matrix form of the second derivative of f_i is

$$(f_i)'' = diag(\frac{\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{i1}+\gamma_k X_{i1}+R_{i1}}}{(1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{i1}+\gamma_k X_{i1}+R_{i1}})^2}, \ldots, \frac{\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{in_i}+\gamma_k X_{in_i}+R_{in_i}}}{(1+\sum_{k=1}^{K-1} e^{\alpha_k+\beta_k G_{in_i}+\gamma_k X_{in_i}+R_{in_i}})^2}) + \frac{(\Sigma_{kin})^{-1}}{\sigma^2} \tag{2.10}$$

Then we apply the coordinate transformation using AGQ to the likelihood resulting in the following log likelihood approximation:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \int e^{-f_i} dR_i = \prod_{i=1}^{n} \int e^{-f_i+\frac{z_i^2}{2}} e^{-\frac{z_i^2}{2}} |(f_i)''|^{-\frac{1}{2}} dz_i \\ &= \prod_{i=1}^{n} (2\pi)^{n_i/2} |(f_i)''|^{-\frac{1}{2}} \int e^{-f_i+\frac{z_i^2}{2}} \frac{1}{(2\pi)^{n_i/2}} e^{-\frac{z_i^2}{2}} dz_i \\ &\approx \prod_{i=1}^{n} (2\pi)^{n_i/2} |(f_i)''|^{-\frac{1}{2}} \sum_{j_1=1}^{N_{GQ}} \cdots \sum_{j_{n_i}=1}^{N_{GQ}} exp\{-f_i + \frac{\left\| z_{j_1}, \ldots, z_{j_{n_i}} \right\|^2}{2}\} \prod_{k=1}^{n_i} w_{jk} \end{aligned} \tag{2.11}$$

Given the approximated log-likelihood function, we estimate the model parameters with the following steps:

1. We preselect a list of possible values of σ^2, for the model containing covariates only (no genetic variant);

2. Given a value of σ^2 and an initial estimate of α, β, γ from the model for unrelated samples, we derive $\hat{R}_i$ that minimizes f_i;

3. Given $\hat{R}_i$ from step 2, we derive $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ that maximizes the AGQ approximated marginal likelihood (2.11);

4. We iterate between steps 2 and 3 until convergence, and calculate the approximated likelihood function.

5. We compare the approximated likelihood function at each σ^2 and select the value of σ^2 with the maximum likelihood function.

2.2.2 Laplace approximation

AGQ is very computer intensive, especially with large number of grid points, although the approximation can be very accurate. We consider the use of Laplace approximation [37] as a more computationally feasible alternative. The coxme package developed by Therneau [44] was also based on Laplace approximation to approximate the survival function with random effects. Once we have the approximated log-likelihood function, we can easily conduct estimation as well as hypothesis testing.

Incorporating the overall family structure, the likelihood function can be rewritten as:

$$L(\theta) = \int f(\theta|R)\phi(R)dR = \frac{1}{(2\pi\sigma^2)^{n/2}|\Sigma_{kin}|^{1/2}} \int e^{logf(\theta|R) - \frac{R'\Sigma_{kin}^{-1}R}{2\sigma^2}} dR \tag{2.12}$$

where $f(\theta|R)$ is the conditional likelihood function; $\phi(R)$ is the density function of R. After rearranging the terms of the intergrand of $L(\theta)$, we define

$$PCl = logf(\theta|R) - \frac{R'\Sigma_{kin}^{-1}R}{2\sigma^2} \tag{2.13}$$

as the ‘‘Penalized Conditional log-likelihood’’ (PCl).

Penalized conditional log-likelihood

The full explicit form of PCl is formulated as

$PCl = log(f(\theta|R)) - \frac{R^T \Sigma_{kin}^{-1} R}{2\sigma^2}$

$= log \prod_{i=1}^{n} \prod_{j=1}^{n_i} \{[\prod_{k=1}^{K-1} (\frac{e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{1+\sum e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}})^{I(Y_{ij}=k)}](\frac{1}{1+\sum e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}})^{I(Y_{ij}=K)}\} - \frac{R^T \Sigma_{kin}^{-1} R}{2\sigma^2}$

$= \sum_{i,j} [\sum_{k=1}^{K-1} I(Y_{ij} = k) log \frac{e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{1+\sum e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}} + I(Y_{ij} = K) log \frac{1}{1+\sum e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}] - \frac{R^T \Sigma_{kin}^{-1} R}{2\sigma^2}$

$= \sum_{i,j} [-log(1 + \sum e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}) + \sum_{k=1}^{K-1} I(Y_{ij} = k)(\alpha_k + \beta_k G_{ij} + \gamma_k^T X_{ij} + R_{ij})] - \frac{R^T \Sigma_{kin}^{-1} R}{2\sigma^2}$

$= -\sum_{i,j} log(1 + \sum e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}) + \sum_{i,j,k} I(Y_{ij} = k)(\alpha_k + \beta_k G_{ij} + \gamma_k^T X_{ij}) + \sum_{i,j} (1 - I(Y_{ij} = K)) R_{ij} - \frac{R^T \Sigma_{kin}^{-1} R}{2\sigma^2}$

After applying a second-order multivariate Taylor expansion to PCl,

$$PCl \approx PCl(\hat{\theta}, \hat{R}) + (\theta - \hat{\theta}, R - \hat{R}) PCl''(\hat{\theta}, \hat{R}) (\theta - \hat{\theta}, R - \hat{R})^T / 2 \tag{2.14}$$

where $\hat{\theta}, \hat{R}$ are the global maxima of PCl, so the term of the first derivative vanishes. Assuming $\hat{\theta}$ is good approximation for $\hat{\theta_{MLE}}$,

$$\begin{aligned} PCl(\hat{\theta_{MLE}}, R) \approx PCl(\hat{\theta}, R) &= PCl(\hat{\theta}, \hat{R}) + (R - \hat{R})^T PCl''(\hat{\theta}, \hat{R})_{RR} (R - \hat{R})/2 \\ &= PCl(\hat{\theta}, \hat{R}) - (R - \hat{R})^T H(\hat{\theta}, \hat{R})_{RR} (R - \hat{R})/2 \end{aligned} \tag{2.15}$$

($H(\cdot,\cdot)$ is the observed information matrix based on PCl.)

Therefore, the approximated likelihood function at the MLE is derived as follows:

$$\begin{aligned} L(\hat{\theta}_{MLE}) &= \frac{1}{(\sqrt{2\pi\sigma^2})^n|\Sigma_{kin}|^{1/2}} \int exp\{PCl(\hat{\theta}_{MLE}, R)\}dR \\ &= \frac{e^{PCl(\hat{\theta},\hat{R})}}{(\sqrt{2\pi\sigma^2})^n|\Sigma_{kin}|^{1/2}} \int e^{-\frac{(R-\hat{R})^T H(\hat{\theta},\hat{R})_{RR}(R-\hat{R})}{2}} dR \\ &= \frac{e^{PCl(\hat{\theta},\hat{R})}}{(\sqrt{2\pi\sigma^2})^n|\Sigma_{kin}|^{1/2}} (\sqrt{2\pi})^n |H|^{-1/2} \\ &= \frac{e^{PCl(\hat{\theta},\hat{R})}}{\sigma^n|\Sigma_{kin}|^{1/2}|H|^{1/2}} \end{aligned} \quad (2.16)$$

Equivalently, the approximated log likelihood function at the MLE is

$$l(\hat{\theta}_{MLE}) = PCl(\hat{\theta}, \hat{R}) - \frac{n}{2}log(\sigma^2) - \frac{1}{2}log|\Sigma_{kin}| - \frac{1}{2}log|H(\hat{\theta}, \hat{R})_{RR}|. \quad (2.17)$$

Grid search

Based on the complex form of the marginal likelihood, we can not solve for the MLE of the variance parameter σ^2 directly. Therefore, similar to the procedure applied in AGQ, we implement a grid-search scheme to obtain the approximated $\hat{\sigma^2}_{MLE}$. The core strategy consists of the following four steps:

1. We choose a set of possible σ^2 values;
2. For a given σ^2, we use Newton-Ralphson algorithm to obtain $\hat{\theta}$ and $\hat{R}$ that maximizes PCl;
3. We calculate the approximated maximum log-likelihood evaluated at the given σ^2;

4. We repeat steps 2 and 3 for all possible value of σ^2 chosen in step 1;

5. We select the value of σ^2 with the largest maximum log-likelihood.

We implement a nested grid search in practice, so that we are able to reduce the searching range and refine the estimates, at each iteration. We first use a large but sparse grid to find the best targeted area, so that in the second search we can use a more dense grid to efficiently and accurately locate the value of $\hat{\sigma^2}_{MLE}$. Here the log likelihood function is assumed to be concave as a function of σ^2, so there exists only one global maxima. The MLE of σ^2 is almost invariant to models with or without the genetic variant. For ease of computation, we implement a nested grid search for σ^2_{MLE} in the model without the genetic variant. Once the best $\hat{\sigma^2}$ is located, we use Newton-Ralphson algorithm to solve for the global maxima $\hat{\theta}$ and $\hat{R}$ of PCl.

Note that in step 2 of the search scheme, θ and R are estimated simultaneously. However, in AGQ, θ and R are estimated recursively. This explains why Laplace approximation improves the computational efficiency of AGQ.

Newton-Ralphson algorithm

The central computation strategy is to use line search as the outer loop and Newton-Ralphson Algorithm as the inner loop. The outerloop searches over the parameter space of σ^2 for the MLE. For a given σ^2, the inner loop works as follows:

1. We solve for the maxima ($\hat{\theta}$, $\hat{R}$) of PCl using Newton-Ralphson Algorithm;

2. We use the formula (2.17) to compute the corresponding maximum log-likelihood.

The general idea of Newton-Raphson method is to utilize a first order approximation assuming the initial starting point does not deviate much from the true value. Generally, θ denotes the vector of parameters to estimate, I denotes the observed information matrix,

and l' denotes the score function. The iteration formula (from the m-th to the (m+1)th iteration) is

$\theta_{(m+1)} = \theta_{(m)} + I(\theta_m)^{-1} l'(\theta_{(m)})$.

The first and second derivative of PCl are key components for the implementation of Newton-Ralphson Algorithm. Let P denote the number of predictors including intercept; $k = 1, \cdots, (K-1)$ and K is the total number of categories. The first-order derivatives of PCl are:

$$\begin{aligned}
\frac{\partial PCl}{\partial \alpha_k} &= -\sum_{i,j} \frac{e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}} + \sum_{i,j} I(Y_{i,j}=k) \\
\frac{\partial PCl}{\partial \beta_k} &= -\sum_{i,j} \frac{G_{ij} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}} + \sum_{i,j} G_{ij} I(Y_{i,j}=k) \\
\frac{\partial PCl}{\partial \gamma_{km}} &= -\sum_{i,j} \frac{X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}} + \sum_{i,j} X_{ijm} I(Y_{i,j}=k) \\
\frac{\partial PCl}{\partial R_{ij}} &= -\frac{\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}} + \sum_{k=1}^{K-1} I(Y_{ij}=k) - \frac{(\Sigma_{kin}^{-1})_{ij,} R}{\sigma^2}
\end{aligned} \tag{2.18}$$

The second-order derivatives are:$\forall m \neq m'$

$$\frac{\partial^2 PCl}{\partial \alpha_k^2} = -\sum_{i,j} \frac{e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}(1+\sum_{t=1;t\neq k}^{t=(K-1)} e^{\alpha_t+\beta_t G_{ij}+\gamma_t^T X_{ij}+R_{ij}})}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}$$

$$\frac{\partial^2 PCl}{\partial \beta_k^2} = -\sum_{i,j} \frac{G_{ij}^2 e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}(1+\sum_{t\neq k} e^{\alpha_t+\beta_t G_{ij}+\gamma_t^T X_{ij}+R_{ij}})}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}$$

$$\frac{\partial^2 PCl}{\partial \gamma_{km}^2} = -\sum_{i,j} \frac{X_{ijm}^2 e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}(1+\sum_{t\neq k} e^{\alpha_t+\beta_t G_{ij}+\gamma_t^T X_{ij}+R_{ij}})}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}$$

$$\frac{\partial^2 PCl}{\partial \gamma_{km}\partial \gamma_{km'}} = -\sum_{i,j} \frac{X_{ijm}X_{ijm'} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}(1+\sum_{t\neq k} e^{\alpha_t+\beta_t G_{ij}+\gamma_t^T X_{ij}+R_{ij}})}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}$$

$$\frac{\partial^2 PCl}{\partial \alpha_k\partial \beta_k} = -\sum_{i,j} \frac{G_{ij} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}(1+\sum_{t\neq k} e^{\alpha_t+\beta_t G_{ij}+\gamma_t^T X_{ij}+R_{ij}})}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}$$

$$\frac{\partial^2 PCl}{\partial \alpha_k\partial \gamma_{km}} = -\sum_{i,j} \frac{X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}(1+\sum_{t\neq k} e^{\alpha_t+\beta_t G_{ij}+\gamma_t^T X_{ij}+R_{ij}})}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}$$

$$\frac{\partial^2 PCl}{\partial \beta_k\partial \gamma_{km}} = -\sum_{i,j} \frac{G_{ij}X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}(1+\sum_{t\neq k} e^{\alpha_t+\beta_t G_{ij}+\gamma_t^T X_{ij}+R_{ij}})}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}$$

$$\frac{\partial^2 PCl}{\partial R_{ij}^2} = -\frac{\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2} - \frac{(\Sigma_{kin}^{-1})_{ij,ij}}{\sigma^2}$$

$$\frac{\partial^2 PCl}{\partial \alpha_k\partial R_{ij}} = -\frac{e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}$$

$$\frac{\partial^2 PCl}{\partial \beta_k\partial R_{ij}} = -\frac{G_{ij} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}$$

$$\frac{\partial^2 PCl}{\partial \gamma_{km}\partial R_{ij}} = -\frac{X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2} \tag{2.19}$$

$\forall k' \neq k$

$$
\begin{aligned}
\frac{\partial^2 PCl}{\partial \alpha_k \partial \alpha_{k'}} &= \sum_{i,j} \frac{e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2} \\
\frac{\partial^2 PCl}{\partial \alpha_k \partial \beta_{k'}} &= \sum_{i,j} \frac{G_{ij} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2} \\
\frac{\partial^2 PCl}{\partial \alpha_k \partial \gamma_{k'm}} &= \sum_{i,j} \frac{X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2} \\
\frac{\partial^2 PCl}{\partial \beta_k \partial \beta_{k'}} &= \sum_{i,j} \frac{G_{ij}^2 e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2} \\
\frac{\partial^2 PCl}{\partial \beta_k \partial \gamma_{k'm}} &= \sum_{i,j} \frac{G_{ij} X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2} \\
\frac{\partial^2 PCl}{\partial \gamma_{km} \partial \gamma_{k'm}} &= \sum_{i,j} \frac{X_{ijm}^2 e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2} \\
\frac{\partial^2 PCl}{\partial \gamma_{km} \partial \gamma_{k'm'}} &= \sum_{i,j} \frac{X_{ijm} X_{ijm'} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}+R_{ij}}}{(1+\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}+R_{ij}})^2}.
\end{aligned}
\tag{2.20}
$$

When the initial starting point is far from the true value, it can make it difficult or even impossible for the algorithm to converge. To solve this issue, a small adjustment is proposed during the iterations [1] [3] [11]:

$\theta_{(k+1)} = \theta_{(k)} + \gamma I(\theta_k)^{-1} l'(\theta_{(k)})$

where $0 < \gamma < 1$ is a prespecified constant.

We select the coefficient estimates of general logit in unrelated samples as the initial value [14]:

$$
g(P(Y_{ij} = k|G_{ij}, R_{ij})) = \alpha_k + \beta_k G_{ij} + X_{ij}^T \gamma_k \tag{2.21}
$$

where $k = 1, \cdots, (K-1)$ and K is the total number of categories. This improves convergence of the inner loop.

2.2.3 Association test

Two types of association tests are performed and compared to evaluate the global null hypothesis $H_0 : \beta_1 = \cdots = \beta_{K-1} = 0$.

Wald test

The Wald test is based on PCl, its global maxima and $\hat{\sigma^2}$.

The expected Information matrix is calculated with respect to R_{ij} $\forall(i,j)$, using the second-order Taylor expansion approximation. Because we know

$$E[R_{ij}] = 0 \tag{2.22}$$

$$Var(R_{ij}) = \sigma^2 \qquad (\Sigma_{kin})_{ij,ij} = \frac{\sigma^2}{2}, \tag{2.23}$$

we can write

$$-E\frac{\partial^2 PCl}{\partial \alpha_k^2}=\sum_{i,j} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}[\frac{1+\sum_{m\neq k}}{(1+\sum)^2}+\frac{(1-4e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}-2\sum\sum_{m\neq k}+\sum^2)\sigma^2}{4(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial \beta_k^2}=\sum_{i,j} G_{ij}^2 e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}[\frac{1+\sum_{m\neq k}}{(1+\sum)^2}+\frac{(1-4e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}-2\sum\sum_{m\neq k}+\sum^2)\sigma^2}{4(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial \gamma_{km}^2}=\sum_{i,j} X_{ijm}^2 e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}[\frac{1+\sum_{t\neq k}}{(1+\sum)^2}+\frac{(1-4e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}-2\sum\sum_{t\neq k}+\sum^2)\sigma^2}{4(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial \gamma_{km}\partial \gamma_{km'}}=\sum_{i,j} X_{ijm}X_{ijm'} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}[\frac{1+\sum_{t\neq k}}{(1+\sum)^2}+\frac{(1-4e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}-2\sum\sum_{t\neq k}+\sum^2)\sigma^2}{4(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial \alpha_k\partial \beta_k}=\sum_{i,j} G_{ij} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}[\frac{1+\sum_{m\neq k}}{(1+\sum)^2}+\frac{(1-4e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}-2\sum\sum_{m\neq k}+\sum^2)\sigma^2}{4(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial \alpha_k\partial \gamma_{km}}=\sum_{i,j} X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}[\frac{1+\sum_{m\neq k}}{(1+\sum)^2}+\frac{(1-4e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}-2\sum\sum_{m\neq k}+\sum^2)\sigma^2}{4(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial \beta_k\partial \gamma_{km}}=\sum_{i,j} G_{ij}X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}[\frac{1+\sum_{m\neq k}}{(1+\sum)^2}+\frac{(1-4e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}-2\sum\sum_{m\neq k}+\sum^2)\sigma^2}{4(1+\sum)^4}] \quad (2.24)$$

$\forall k \neq k'$

$$-E\frac{\partial^2 PCl}{\partial\alpha_k\partial\alpha_{k'}}=-\sum_{i,j} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}}[\frac{1}{(1+\sum)^2}+\frac{(2-\sum)\sigma^2}{2(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial\beta_k\partial\beta_{k'}}=-\sum_{i,j} G_{ij}^2 e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}}[\frac{1}{(1+\sum)^2}+\frac{(2-\sum)\sigma^2}{2(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial\gamma_{km}\partial\gamma_{k'm}}=-\sum_{i,j} X_{ijm}^2 e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}}[\frac{1}{(1+\sum)^2}+\frac{(2-\sum)\sigma^2}{2(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial\alpha_k\partial\beta_{k'}}=-\sum_{i,j} G_{ij} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}}[\frac{1}{(1+\sum)^2}+\frac{(2-\sum)\sigma^2}{2(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial\alpha_k\partial\gamma_{k'm}}=-\sum_{i,j} X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}}[\frac{1}{(1+\sum)^2}+\frac{(2-\sum)\sigma^2}{2(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial\beta_k\partial\gamma_{k'm}}=-\sum_{i,j} G_{ij} X_{ijm} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}}[\frac{1}{(1+\sum)^2}+\frac{(2-\sum)\sigma^2}{2(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial\gamma_{km}\partial\gamma_{k'm'}}=-\sum_{i,j} X_{ijm} X_{ijm'} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}} e^{\alpha_{k'}+\beta_{k'} G_{ij}+\gamma_{k'}^T X_{ij}}[\frac{1}{(1+\sum)^2}+\frac{(2-\sum)\sigma^2}{2(1+\sum)^4}]$$

$$-E\frac{\partial^2 PCl}{\partial R_{ij}^2}=\sum e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}[\frac{1}{(1+\sum)^2}+\frac{(1-4\sum+\sum^2)\sigma^2}{4(1+\sum)^4}] \tag{2.25}$$

where $\sum$ is the abbreviation for $\sum_{k=1}^{k=(K-1)} e^{\alpha_k+\beta_k G_{ij}+\gamma_k^T X_{ij}}$.

Let θ denote the coefficient parameters $\theta = (\alpha, \beta, \gamma, R)$. Because R is a nuissance vector,

we adjust the information matrix by eliminating the impact of R [46]:

$$I_{\theta\theta adj} = I_{\theta\theta} - I_{\theta R} I_{RR}^{-1} I_{R\theta}. \tag{2.26}$$

The Wald test statistic is

$$\chi^2 = \hat{\beta}^T ((I_{\theta\theta adj}^{-1})_{\beta\beta})^{-1} \hat{\beta} \tag{2.27}$$

evaluated at the global maxima $\hat{\theta^2}$ of PCl and follows a χ^2_{K-1} distribution asymptotically.

Likelihood ratio test

The likelihood ratio statistic is defined as

$$\chi^2 = -2(l_0 - l_a), \tag{2.28}$$

where l_0 is the maximum log likelihood evaluated at the null hypothesis, and l_a is the maximum log likelihood evaluated at the whole space (union of the null and the alternative hypotheses) of the parameters. From statistics theory, we know the test statistic follows a $\chi^2_{(K-1)}$ distribution asymptotically.

2.3 Simulation studies

We perform simulation studies to assess the type-I error rate and the power under several different scenarios.

2.3.1 Type-I error assessment

Five thousand datasets are simulated to assess the type-I error of the proposed model. In each dataset, 500 nuclear families with 2 offspring are simulated. Two continuous phenotypes with the same moderate heritability ($h^2 \approx 0.43$), FG and BMI, are generated. We use these two continuous traits to create a three-class multinomial

trait (diabetic & obese, diabetic & non-obese , non-diabetic) which has no obvious ordinal trend. An additively coded genotype is generated using random dropping within families, independently of the phenotypes. The variance of the bivariate phenotypes in the i^{th} family is

$$\begin{aligned} Var(Y_{i\cdot}) &= \Sigma_{G_i} + \Sigma_{E_i} = \Sigma_A \otimes (2 \times \Sigma_{kin_i}) + \Sigma_E \otimes I_{n_i} \\ &= \begin{pmatrix} var(FG) & cov(FG, BMI) \\ cov(FG, BMI) & var(BMI) \end{pmatrix} \otimes (2 \times \Sigma_{kin_i}) + \Sigma_E \otimes I_{n_i}. \end{aligned} \tag{2.29}$$

We denote the heritability of the continuous trait as h^2, which is the proportion of variance in the trait explained by additive genetic effect. Thus

$$\Sigma_E = (\frac{1}{h^2} - 1)\Sigma_A \tag{2.30}$$

assuming the same heritability for both traits, where $\Sigma_A = \begin{pmatrix} var(FG) & cov(FG, BMI) \\ cov(FG, BMI) & var(BMI) \end{pmatrix}$ is estimated from FHS data. 2 cutoffs of discretization to generate the 3 categories (i.e., diabetic $\&$ obese, diabetic $\&$ non-obese and non-diabetic, Table 2.1) are explored. We use $\alpha = 0.05$ to declare statistical significance. To summarize, a total of 4 scenarios are investigated:

1. Genotype with MAF=0.3, a trait with three even categories;

2. Genotype with MAF=0.05, a trait with three even categories;

3. Genotype with MAF=0.3, a trait with three uneven categories (22%, 18%, 60%);

4. Genotype with MAF=0.05, a trait with three uneven categories (22%, 18%, 60%);

Table 2.1: Distribution of evenly/unevenly distributed Categories

Category	1. diabetic & obese	2. diabetic & non-obese	3. non-diabetic
Uneven	22%	18%	60%
Even	33.3%	33.3%	33.3%

Table 2.2: Type-I error results

α	MAF	Design	Wald (95% CI)	LRT (95% CI)	GLMM (LRT)
0.05	0.3	Even	0.048 (0.043, 0.055)	0.060 (0.054, 0.067)	0.0484
		Uneven	0.054 (0.048, 0.061)	0.065 (0.059, 0.072)	0.0504
	0.05	Even	0.048 (0.042, 0.055)	0.062 (0.056, 0.069)	0.0458
		Uneven	0.048 (0.042, 0.054)	0.058 (0.051, 0.064)	0.043

We caculate the type-I error rate for our approach using both wald-test and LRT (Table 2.2), and chose wald-test as the association test for our approach, because it has the correct type-I error rate compared to the slight inflation of LRT.

2.3.2 Power assessment

Power of this proposed model is compared to 3 other approaches. 5000 datasets are simulated under the alternative hypothesis that the phenotype is associated with the genotype. In each replicate, 300 nuclear families with 4 offspring are generated. The simulated genotype accounts for 0.5% of the total trait variability. Three categories are assigned based on the following probability equations.

$$\begin{cases} log\frac{P(Y_{ij}=1)}{P(Y_{ij}=3)} = \beta_{10} + \beta_1 SNP_{ij} + R_{ij} \\ log\frac{P(Y_{ij}=2)}{P(Y_{ij}=3)} = \beta_{20} + \beta_2 SNP_{ij} + R_{ij} \end{cases}$$

where

β_{10} and β_{20} are randomly selected from a uniform distribution between -0.1 and 0.1 ($U(-0.1, 0.1)$)

$max(|\beta_1|, |\beta_2|) = \sqrt{\frac{0.5\%}{2MAF(1-MAF)}}$

$min(|\beta_1|, |\beta_2|) = \frac{1}{3}\sqrt{\frac{0.5\%}{2MAF(1-MAF)}}$. The sign of β_1 and β_2 are randomly assigned with equal probabilities.

The vector R is generated from a multivariate normal distribution, $R \sim N(0, \sigma_a^2\Sigma_{kin})$ $(0 < \sigma_a^2 < 1)$. Each individual is assigned to the category with the highest probability (i.e. $arg_k max(P(Y_{ij} = k))$, based on his genotype and the above equations.

The three methods against which we compare our proposed model (Fammulti) are Generalized Linear Mixed Model (GLMM) clustered by family, using SAS procedure GLIMMIX and IML; our proposed model with a binary outcome by collapsing categories 2 and 3; and GLMM clustered by families with categories 2 and 3 collapsed. Two genetic variants, one with MAF=0.05 and the other with MAF=0.3, are studied. Two thresholds (0.01 and 0.001) are used to claim a significance. Our approach is consistently more powerful than the other three methods (Table 2.3). The type-I error rate of GLMM has been explored and demonstrated to be correct (Table 2.2), so we are comparing the power of our approach with GLMM on a fair basis.

Table 2.3: Power rate of Fammulti, GLMM, Collapsed Fammulti and Collapsed GLMM

MAF	α	Fammulti	GLMM	Collapsed Fammulti	Collapsed GLMM
0.3	0.01	77.9%	72.6%	47.2%	2.6%
	0.001	63.3%	54.6%	33.8%	1.4%
0.05	0.01	80.1%	77.1%	50.5%	1.7%
	0.001	64.2%	58.1%	36.2%	$< 1\%$

2.4 Application

2.4.1 Phenotype

Our novel approach is applied to a real phenotype dataset from the FHS with a sample size of 5709 with no missing phenotype or covariates. Initiated in 1948, the FHS is a

longitudinal study consisting of three generations of participants: the original cohort, the offspring cohort and the 3rd generation cohort, comprising 14428 participants. All the participants are from the town of Framingham, Massachusetts, and some participants are related. Over the years, FHS has yielded fruitful results in identifying risk factors of cardiovascular-related traits like blood pressure, cholesterol level as well as glycemic and metabolic traits. Moreover, the association between the physical traits and genetic factors are also being studied intensively. We first create an obesity trait by categorizing the value of BMI at exam 5 in the Offspring cohort participants and the first exam in the Generation 3 cohort participants. Although obesity usually has four categories: normal ($BMI \leq 25$); overweight ($25 \leq BMI \leq 30$); moderately obese ($30 \leq BMI \leq 35$) and severely obese ($BMI > 35$), we collapse the upper two categories (moderately obese and severely obese) because each of them includes few individuals.The proportion in each category is presented in Table 2.4. Because BMI increases with age, on average, we adjust our analyses for age of the participants, which ranges from 19 to 82 years.

Table 2.4: Proportion of various obesity statuses in the phenotype dataset

Normal	Overweight	Moderately or severely obese
38.6%	38.3%	23.1%

2.4.2 Genotype

We apply our approach to MACH (Markov Chain based haplotyper)-imputed SHARe genotype data using Affymetrix 500K array supplemented by the MIPH 50K array. The NHLBI SHARe Project conducted genome wide association studies (GWAS) in several large, multi-ethnic NHLBI Cohort studies of men and women to identify genes underlying cardiovascular and lung disease and other disorders such as osteoporosis and diabetes. The Framingham SHARe was the first cohort released in Oct 2007, with genotypes from 550,000 SNPs in over 9200 participants. Additional SNPs were imputed in the software

MACH [23] with HapMap 2 reference haplotypes developed by researchers from the University of Michigan. Prior GWAS has indicated that a gene on chromosome 16, *FTO*, was associated with BMI level, so we perform a chromosome-wide association test with obesity status on chromosome 16, to evaluate the feasibility of our approach in a large cohort study.

2.4.3 Results

We show the overall results visually by means of the chromosome-wide plot. We observe all the top SNPs are on four genes: *CDH13* (82.66-83.83 Mb) , *FTO* (53.74 – 54.16 Mb), *PKD1L2* (81.13-81.25 Mb) and *WWOX* (78.13 – 79.25 Mb). *CDH13* is known to be associated with plasma levels of adiponectin [21] [8] [30] [7] [50], a trait correlated with BMI [33]. The strongest *CDH13* signal is rs4782798 ($p = 9.7 \times 10^{-6}$). *FTO* is a known obesity gene [12], with the strongest signal observed at rs1558902 ($p = 10^{-4}$). *PKD1L2* a kidney-disease associated gene also known to be implicated in Basal Metabolic Rate (BMR), with the strongest signal observed at rs9921509 ($p = 10^{-4}$). *WWOX* is known to be implicated in both HDL cholesterol [38] and hypertension [51]. The strongest WWOX signal is located at rs2667621 (p=7.1×10^{-5}).

2.5 Discussion

In this chapter, we propose a novel approach to test the association between a genetic variant and multinomial phenotypes in family samples. We use Laplace method to approximate the marginal likelihood, and to calculate the MLE of the variance component. By combining the Newton-Ralphson Algorithm with the Laplace approximation, our approach is flexible and computationally efficient for medium to large pedigrees.

We recommend using Wald test for both common and rare variants, because our

Table 2.5: Top 30 SNPs on Chromosome 16

Gene	rsid	Position	p
CDH13	rs4782798	83505688	9.7×10^{-6}
	rs16960609	83510664	1.2×10^{-5}
	rs9925903	83508437	1.2×10^{-5}
	rs2325834	83499954	1.5×10^{-5}
	rs4782797	83498540	1.6×10^{-5}
	rs8050667	83498650	1.6×10^{-5}
	rs10871273	83504445	1.6×10^{-5}
	rs12597141	83503576	1.6×10^{-5}
	rs2325831	83500158	1.6×10^{-5}
	rs11860430	83499166	1.7×10^{-5}
	rs2054917	83503237	1.9×10^{-5}
	rs8061757	83516468	2.7×10^{-5}
	rs1873142	83500251	3.0×10^{-5}
	rs7188893	83519289	3.3×10^{-5}
intergenic	rs990346	26797320	3.9×10^{-5}
CDH13	rs17685702	83509734	5.1×10^{-5}
WWOX	rs2667621	78566150	7.1×10^{-5}
CDH13	rs4782796	83491379	7.3×10^{-5}
	rs4782795	83491266	7.3×10^{-5}
	rs7185723	83492640	9.0×10^{-5}
	rs1387381	83492734	9.0×10^{-5}
	rs11149571	83494888	9.0×10^{-5}
	rs9930243	83514087	9.0×10^{-5}
	rs7187676	83492854	9.0×10^{-5}
	rs1552557	83493294	9.1×10^{-5}
	rs7194439	83493961	9.1×10^{-5}
	rs6563910	83494399	9.5×10^{-5}
	rs12716737	83492164	9.7×10^{-5}
PKD1L2	rs9921509	81192256	0.000107
CDH13	rs12445788	83495789	0.000121
FTO	rs1558902	53803574	1.3×10^{-4}

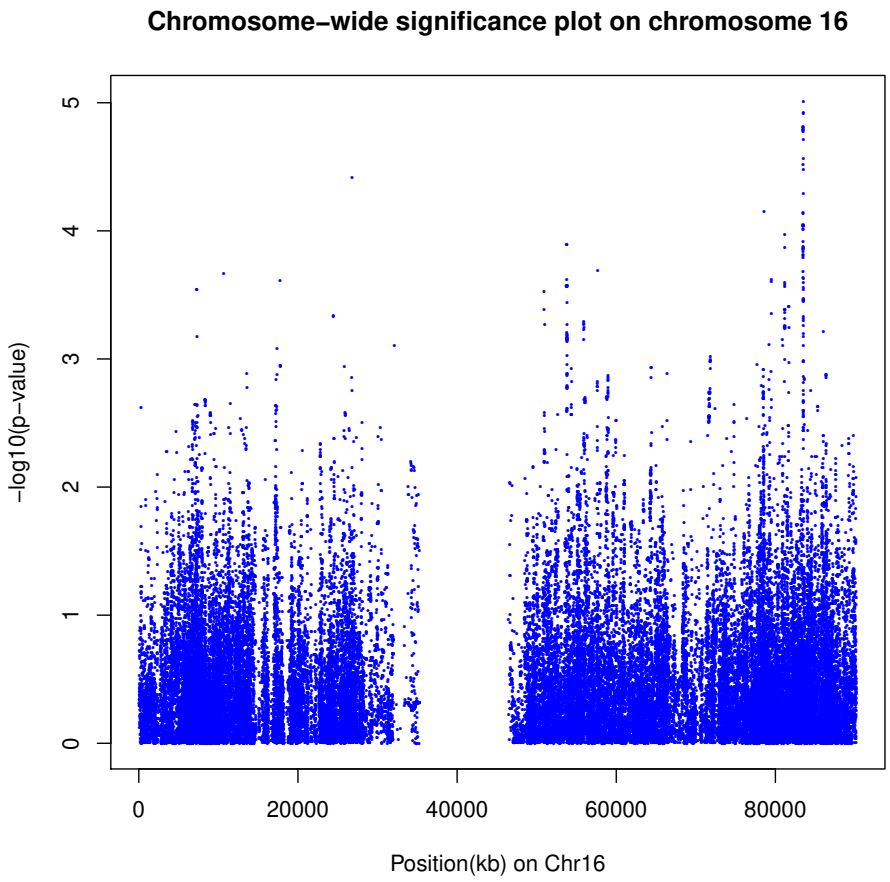

Figure 2-1: Association results for multi-category obesity status and SNPs on chromosome 16

simulation studies demonstrate that the Wald test always gives the correct type-I error rate, while LRT has slightly inflated type-I error rate. We postulate the reason behind is: the Wald test statistic uses the true MLE of PCl, while LRT statistic is approximated by Laplace approximation, and hence there could be more bias in the LRT statistic than the Wald test statistic.

Compared to GLMM, our approach is consistently more powerful, for all the MAF scenarios studied. Our approach has a few advantages over GLMM. First, our approach takes into consideration the pairwise kinship distance instead of assuming a correlation structure that doesn't take relationship into consideration as in GLMM. This is essential, because it is not always possible to have relevant information to decide which covariance structure is the best to use for GLMM. Second, GLMM can be very underpowered, especially in the scenario when some of the categories are collapsed.

Our approach has the potential to discover novel associated genes. In the data analysis, in addition to the obesity gene *FTO*, we also identify variants on the adiponectin gene *CDH13* to be strongly ($p \sim 10^{-5}$) associated with the multinomial obesity status. Similarly, we identify genes associated with other metabolic traits like HDL cholesterol, hypertension and BMR. These results are not contradictory, because some of the metabolic traits like HDL cholesterol and BMR are strongly associated with BMI [36] [28]. In this chapter, we study the multinomial model assuming it has a canonical link function. However, other link functions sometimes may fit the data better and may facilitate the interpretation in certain conditions. So in the future, we would work on exploring the multinomial model with a general applicable link function.

Chapter 3

Bivariate association analysis with Extended Generalized Estimating Equations in family samples

3.1 Introduction

Univariate association test has been widely used in genetics epidemiology and, when applied to GWAS, has yielded fruitful results. However, for correlated phenotypes, the univariate association tests are not as powerful or efficient as multivariate tests. In the case of two continuous phenotypes assumed to be normally distributed, a joint test can be derived as a simple extension of a univariate normal test. However, if one of the two traits is a discrete trait, for example, a binary trait, it is more challenging to derive a multivariate test of association, and it becomes even more challenging in family samples. One reason is that there is no closed form of the likelihood function for a binary trait in family samples. Quasi-likelihood-based approaches have been proposed to address such questions, and the mostly widely known approach is the GEE [54]. GEE has been frequently used to analyze correlated data and perform univariate association tests. Hall and Severini extended GEE

to EGEE [16] in 1998. EGEE has proved to be more powerful and more efficient than GEE while retaining many of the good properties of GEE. Liu et al. [25] successfully implemented EGEE in the context of a joint association test of continuous and binary traits in unrelated samples. Here we propose a method to conduct association tests for bivariate phenotypes in family samples based on EGEE.

3.2 Methods

We first define the model equations for the two phenotypes in family samples, as well as the notations and the assumptions. We assume that there are N independent families ($i = 1, \cdots, N$), and the family size (n_i) depends on the family index (i). The model is composed of two simultaneous equations written as:

$$\begin{cases} Y_{ija} = X_{ij}^T \beta_1 + b_{01ij} + \epsilon_{ij} \\ g(\mu_{ijb}) = g(E[Y_{ijb}]) = X_{ij}^T \beta_2 + b_{02ij} \end{cases}$$

where

i is the family index, while j ($j = 1, ..., n_i$) represents the j-th individual in the i-th family;

Y_{ija} is a quantitative trait, and Y_{ijb} is a binary trait;

Σ_{kin} is the kinship matrix derived from the overall pedigree;

$b_{01} = \begin{pmatrix} b_{0111} \\ \vdots \\ b_{01Nn_N} \end{pmatrix} \sim N(0, \Sigma_{kin}\sigma_1^2)$ is the random vector to account for the familial correlation of the quantitative trait, and

$b_{02} = \begin{pmatrix} b_{0211} \\ \vdots \\ b_{02Nn_N} \end{pmatrix} \sim N(0, \Sigma_{kin}\sigma_2^2)$ is the random vector to account for the familial correlation of the binary trait. The two familial vectors b_{01} and b_{02} are assumed to be independent.

The term $\epsilon_{ij} \sim N(0, \sigma_e^2)$ is the random error term of the quantitative trait.

We define the overall variance matrix of the bivariate phenotypes as $V = diag\{V_1, \cdots, V_N\}$, where V_i $(i = 1, ..., N)$ is the variance matrix of the bivariate phenotypes for the ith family with a dimension of $2n_i \times 2n_i$. It has a form of $\begin{pmatrix} Var(Y_{ia}) & cov(Y_{ia}, Y_{ib}) \\ cov(Y_{ib}, Y_{ia}) & Var(Y_{ib}) \end{pmatrix}$. Because the variance matrix is a vital component of estimating the parameters as well as conducting hypothesis testing, we derive all the elements of the variance matrix in details in 3.2.1.

3.2.1 Variance structure

1. $Var(Y_{ia})$ is the covariance matrix of the quantitative trait for the ith family, and is defined as

$$Var(Y_{ia}) = \begin{pmatrix} var(Y_{i1a}) & \cdots & cov(Y_{i1a}, Y_{ija}) & \cdots & cov(Y_{i1a}, Y_{in_ia}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(Y_{ija}, Y_{i1a}) & \cdots & var(Y_{ija}) & \cdots & cov(Y_{ija}, Y_{in_ia}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(Y_{in_ia}, Y_{i1a}) & \cdots & cov(Y_{in_ia}, Y_{ija}) & \cdots & var(Y_{in_ia}) \end{pmatrix};$$

2. $cov(Y_{ia}, Y_{ib})$ is the covariance matrix of the quantitative trait and the binary trait, and is defined as

$$cov(Y_{ia}, Y_{ib}) = \begin{pmatrix} cov(Y_{i1a}, Y_{i1b}) & \cdots & cov(Y_{i1a}, Y_{ijb}) & \cdots & cov(Y_{i1a}, Y_{in_ib}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(Y_{ija}, Y_{i1b}) & \cdots & cov(Y_{ija}, Y_{ijb}) & \cdots & cov(Y_{ija}, Y_{in_ib}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(Y_{in_ia}, Y_{i1b}) & \cdots & cov(Y_{in_ia}, Y_{ijb}) & \cdots & cov(Y_{in_ia}, Y_{in_ib}) \end{pmatrix};$$

3. $cov(Y_{ib}, Y_{ia})$ is the covariance matrix of the binary trait and the quantitative trait, and is defined as

$$cov(Y_{ib}, Y_{ia}) = \begin{pmatrix} cov(Y_{i1b}, Y_{i1a}) & \cdots & cov(Y_{i1b}, Y_{ija}) & \cdots & cov(Y_{i1b}, Y_{in_ia}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(Y_{ijb}, Y_{i1a}) & \cdots & cov(Y_{ijb}, Y_{ija}) & \cdots & cov(Y_{ijb}, Y_{in_ia}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(Y_{in_ib}, Y_{i1a}) & \cdots & cov(Y_{in_ib}, Y_{ija}) & \cdots & cov(Y_{in_ib}, Y_{in_ia}) \end{pmatrix}$$
$$= cov(Y_{ia}, Y_{ib})^T;$$

4. $Var(Y_{ib})$ is the covariance matrix of the binary trait for the ith family, and is defined as

$$Var(Y_{ib}) = \begin{pmatrix} var(Y_{i1b}) & \cdots & cov(Y_{i1b}, Y_{ijb}) & \cdots & cov(Y_{i1b}, Y_{in_ib}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(Y_{ijb}, Y_{i1b}) & \cdots & var(Y_{ijb}) & \cdots & cov(Y_{ijb}, Y_{in_ib}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(Y_{in_ib}, Y_{i1b}) & \cdots & cov(Y_{in_ib}, Y_{ijb}) & \cdots & var(Y_{in_ib}) \end{pmatrix};$$

All the elements of the variance matrix are calculated using second-order Taylor expansion with respect to $b = (b_{01}, b_{02})$. For derivation details, please see the Appendix.

3.2.2 Conditional correlation matrix

The r, $r_{jj'}$ used in the above variance calculations (see the Appendix) are the correlation parameters which measure the correlation between the two types of traits for the same individual, and any two individuals (j, j'). Generally, we can construct a pedigree-based correlation matrix based on the kinship matrix obtained from a pedigree file. There are a few ways to specify the conditional correlation matrix. However, among the different options, one consensus is the conditional correlation is 0 between any unrelated pairs. e.g., founders, non-inbreed couples, and any two individuals from two independent families.

The more distant the biological relations of the two individuals, the smaller the conditional correlation is. Since they are correlation parameters, they are supposed to be between -1 and 1. Here I propose two feasible options:

1. Two parameters (r, ρ):

$$r_{jj'} = \begin{cases} r & (\Sigma_{kin})_{jj'} = 0.5 \\ r^{\rho} & 0 < (\Sigma_{kin})_{jj'} < 0.5 \\ 0 & (\Sigma_{kin})_{jj'} = 0 \end{cases} \tag{3.1}$$

2. One parameter (r):

$$r_{jj'} = \begin{cases} r^{1/(2\times(\Sigma_{kin})_{jj'})} & (\Sigma_{kin})_{jj'} \neq 0 \\ 0 & (\Sigma_{kin})_{jj'} = 0 \end{cases} \tag{3.2}$$

The first parameterization is constrained by two parameters (r, ρ), which is similar to the compound symmetry covariance structure used in the LME, except that the elements for any unrelated pairs is set to 0. While the second has only one parameter r and therefore works more generally and leaves more degrees of freedom.

3.3 Quasi-likelihood

Quasi-likelihood mimics but is not a real likelihood function. However, it does have many of the good properties of a likelihood function. For example, the most basic form of quasi-likelihood can be constructed as

$$Q(\mu; y) = \int_{y}^{\mu} \frac{y - t}{\sigma^2 V(t)} dt. \tag{3.3}$$

The quasi-likelihod equation in 3.3 behaves like a likelihood function, because $U = \frac{\partial Q(\mu;y)}{\partial \mu} = \frac{Y-\mu}{\sigma^2 V(\mu)}$ satisfies the major properties of a score function:

1. $E(U) = 0$;
2. $Var(U) = \frac{1}{\sigma^2 V(\mu)}$;
3. $-E(\frac{\partial U}{\partial \mu}) = \frac{1}{\sigma^2 V(\mu)}$.

A more stringent condition requires that the score function of the quasi-likelihood to be the gradient factor, like the score function of a real likelihood function, satisfying one additional condition:

$$\frac{\partial^2 U(\beta)}{\partial \beta_r \beta_s} = \frac{\partial^2 U(\beta)}{\partial \beta_s \beta_r} \qquad \forall (r \neq s). \tag{3.4}$$

However, this condition generally does not hold. After reparameterization the quasi-likelihood function has the following general form (assuming no dispersion) [29]:

$$Q(\mu; y) = -(y-\mu)^T \int_0^1 s[V(t(s))]^{-1} ds (y-\mu), \tag{3.5}$$

where $t(s) = y + (\mu - y)s$

3.4 Incorporating the correlation information into the quasi-likelihood

The typical quasi-likelihood function does not allow for correlation parameters, explaining why the correlation parameters are not estimable in GEE. Hall and Severni improved the quasi-likelihood function by incorporating the correlation parameters, such that

$$Q^+(\mu; y) = Q(\mu; y) + f_1(a) + f_2(y) \tag{3.6}$$

where a is a vector containing s correlation parameters. According to quasi-likelihood theory, in order to remain a quasi-likelihood function, the following condition needs to be satisfied: $\forall(u = 1, \cdots, s)$

$$\frac{\partial Q^{+}(\mu; y)}{\partial a_u} = 0. \tag{3.7}$$

After solving these constraint equations, we write the extended quasi-likelihood function as

$$Q^{+}(\mu, a; y) = -\sum_{i=1}^{i=K}(y_i - \mu_i)^T \int_0^1 s[V_i(t(s))]^{-1} ds(y_i - \mu_i) + \frac{1}{2} log|V^{-1}|. \tag{3.8}$$

3.5 Parameter estimation

The parameter estimation is based on solving the Score Equations which converges in just a few iterations. For computer efficiency, we adopt the convenient form of the score equations [15] written as:

$$\begin{aligned}\sum_{i=1}^{i=N} U_i(\beta, \tilde{\alpha}) &= \sum_{i=1}^{i=N} \begin{pmatrix} D_i' & 0 \\ 0 & F_i' \end{pmatrix} \begin{pmatrix} V_i^{-1} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{pmatrix} \\ &= \sum_{i=1}^{i=N} \begin{pmatrix} D_i' V_i^{-1}(y_i - \mu_i) \\ F_i'(s_i - \sigma_i) \end{pmatrix} = 0\end{aligned} \tag{3.9}$$

where

$D_i = \frac{\partial \mu_i}{\partial \beta'}$ is a $2n_i \times (p_1 + p_2)$ matrix (p_1, p_2 is the number of predictors of the continuous

and binary traits respectively). It has two diagonal matrices, and is defined as

$$D_i = \begin{pmatrix} D_{i1} & 0 \\ 0 & D_{i2} \end{pmatrix} = \begin{pmatrix} X_{i1} & 0 \\ 0 & Diag\{\frac{e^{\eta_{i2}}}{(1+e^{\eta_{i2}})^2}\}X_{i2} \end{pmatrix} \tag{3.10}$$

where $\eta_{i2} = (\eta_{i12}, ..., \eta_{in_i2})$. η_{i2} is the linear predictor of the binary trait of the i-th family, and defined as (for $j = 1, ..., n_i$)

$$\eta_{ij2} = x_{ij2}^T \beta_2 \tag{3.11}$$

$F_i = \frac{\partial vec V_i^{-1}}{\partial \tilde{\alpha}'}$ is a $4n_i^2 \times 4$ matrix ($\tilde{\alpha} = (\sigma_1^2, \sigma_2^2, \sigma_e^2, r)$ is the correlation vector)
$F_{i\alpha_k} = vec(-V_i^{-1} \frac{\partial V_i}{\partial \alpha_k} V_i^{-1})$ ($\alpha_1 = \sigma_1^2$, $\alpha_2 = \sigma_2^2$, $\alpha_3 = \sigma_e^2$, $\alpha_4 = r$)

$$\frac{\partial V_i}{\partial \sigma_1^2} = \begin{pmatrix} \Sigma_{kin}^i & 0 \\ 0 & 0 \end{pmatrix} \tag{3.12}$$

$$\frac{\partial V_i}{\partial \sigma_2^2} = \begin{pmatrix} 0 & W_1 \\ W_1^T & W_2 \end{pmatrix} \tag{3.13}$$

$$W_1 = \frac{\sigma_e}{16} Rdiag\{\frac{e^{X_{ij}^T\beta_2/2}(1 + e^{X_{ij}^T\beta_2})(1 - 6e^{X_{ij}^T\beta_2} + e^{2X_{ij}^T\beta_2})}{\left(1 + e^{X_{ij}^T\beta_2}\right)^4}\}. \tag{3.14}$$

The elements of W_2 are $\forall\ j \neq j'$

$$w_{2jj} = \frac{\partial var(Y_{ijb})}{\partial \sigma_2^2} = \frac{e^{X_{ij}^T\beta_2}\left(1-e^{X_{ij}^T\beta_2}\right)^2}{4\left(1+e^{X_{ij}^T\beta_2}\right)^4} - \frac{e^{2X_{ij}^T\beta_2}\left(1-e^{X_{ij}^T\beta_2}\right)^2\sigma_2^2}{8\left(1+e^{X_{ij}^T\beta_2}\right)^6}$$

$$w_{2jj'}=\frac{\partial cov(Y_{ijb},Y_{ij'b})}{\partial \sigma_2^2}=\frac{e^{(X_{ij}+X_{ij'})^T\beta_2}}{\left(1+e^{X_{ij}^T\beta_2}\right)^2\left(1+e^{X_{ij'}^T\beta_2}\right)^2}\left((\Sigma_{kin})_{ij,ij'}-\frac{(1-e^{X_{ij}^T\beta_2})(1-e^{X_{ij'}^T\beta_2})\sigma_2^2(\Sigma_{kin})_{ij,ij}(\Sigma_{kin})_{ij',ij'}}{2(1+e^{X_{ij}^T\beta_2})(1+e^{X_{ij'}^T\beta_2})}\right)$$

(3.15)

$$\frac{\partial V_i}{\partial \sigma_e^2} = \begin{pmatrix} 0 & W_3 \\ W_3^T & 0 \end{pmatrix} \tag{3.16}$$

$$W_3 = \frac{1}{2\sqrt{\sigma_e^2}} Rdiag\{\left(\frac{e^{X_{ij}^T\beta_2/2}}{1+e^{X_{ij}^T\beta_2}} + \frac{1}{16}\frac{e^{X_{ij}^T\beta_2/2}(1+e^{X_{ij}^T\beta_2})(1-6e^{X_{ij}^T\beta_2}+e^{2X_{ij}^T\beta_2})}{\left(1+e^{X_{ij}^T\beta_2}\right)^4}\sigma_2^2\right)\} \tag{3.17}$$

$$\frac{\partial V_i}{\partial r} = \begin{pmatrix} 0 & W_4 \\ W_4^T & 0 \end{pmatrix} \tag{3.18}$$

$$W_4 = \frac{\partial R}{\partial r}\sigma_e diag\{\left(\frac{e^{X_{ij}^T\beta_2/2}}{1+e^{X_{ij}^T\beta_2}} + \frac{1}{16}\frac{e^{X_{ij}^T\beta_2/2}(1+e^{X_{ij}^T\beta_2})(1-6e^{X_{ij}^T\beta_2}+e^{2X_{ij}^T\beta_2})}{\left(1+e^{X_{ij}^T\beta_2}\right)^4}\sigma_2^2\right)\} \tag{3.19}$$

$s_i = vec\left\{(y_i-\mu_i)(y_i-\mu_i)'\right\}$

$\sigma_i = Es_i = vecV_i.$

Fisher's scoring algorithm is implemented to solve the score equations iteratively. The updating equation (from the m-th to the (m+1)-th iteration) is :

$$\begin{pmatrix} \beta^{(m+1)} \\ \tilde{\alpha}^{(m+1)} \end{pmatrix} = \begin{pmatrix} \beta^{(m)} \\ \tilde{\alpha}^{(m)} \end{pmatrix} + \left(U^*(\beta^{(m)}, \tilde{\alpha}^{(m)})\right)^{-1} \sum_{i=1}^{i=N} U_i(\beta^{(m)}, \tilde{\alpha}^{(m)}) \tag{3.20}$$

where

$$\begin{aligned} U^*(\beta, \tilde{\alpha}) = -ED\left(\sum_{i=1}^{i=N} U_i(\beta, \tilde{\alpha})\right) &= \sum_{i=1}^{i=N} \begin{pmatrix} D_i' & 0 \\ 0 & F_i' \end{pmatrix} \begin{pmatrix} V_i^{-1} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} D_i & 0 \\ \frac{\partial \sigma_i}{\partial \beta'} & \frac{\partial \sigma_i}{\partial \tilde{\alpha}'} \end{pmatrix} \\ &= \sum_{i=1}^{i=N} \begin{pmatrix} D_i' V_i^{-1} D_i & 0 \\ F_i' \frac{\partial \sigma_i}{\partial \beta'} & F_i' \frac{\partial \sigma_i}{\partial \tilde{\alpha}'} \end{pmatrix} \end{aligned} \tag{3.21}$$

$$\frac{\partial \sigma_i}{\partial \beta_1} = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \tag{3.22}$$

$$\frac{\partial \sigma_i}{\partial \beta_{2p}} = vec \begin{pmatrix} 0 & W_6 \\ W_6^T & W_7 \end{pmatrix} \tag{3.23}$$

$$W_6 = \frac{\sigma_e}{16} Rdiag\{\frac{e^{\frac{X_{ij}^T \beta_2}{2}} X_{ijp}(1 - e^{X_{ij}^T \beta_2})}{(1 + e^{X_{ij}^T \beta_2})^4} \left[8(1 + e^{X_{ij}^T \beta_2})^2 + \frac{1}{2}\sigma_2^2(1 - 22e^{X_{ij}^T \beta_2} + e^{2X_{ij}^T \beta_2})\right]\}. \tag{3.24}$$

W_7 is defined as

$$W_7 = \begin{pmatrix} \frac{\partial var(Y_{i1b})}{\partial \beta_{2p}} & \cdots & \frac{\partial cov(Y_{i1b},Y_{ijb})}{\partial \beta_{2p}} & \cdots & \frac{\partial cov(Y_{i1b},Y_{in_ib})}{\partial \beta_{2p}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial cov(Y_{ijb},Y_{i1b})}{\partial \beta_{2p}} & \cdots & \frac{\partial var(Y_{ijb})}{\partial \beta_{2p}} & \cdots & \frac{\partial cov(Y_{ijb},Y_{in_ib})}{\partial \beta_{2p}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial cov(Y_{in_ib},Y_{i1b})}{\partial \beta_{2p}} & \cdots & \frac{\partial cov(Y_{in_ib},Y_{ijb})}{\partial \beta_{2p}} & \cdots & \frac{\partial var(Y_{in_ib})}{\partial \beta_{2p}} \end{pmatrix} \tag{3.25}$$

and the elements are derived as:

$$\frac{\partial var(Y_{ijb})}{\partial \beta_p} = (1-2E\mu_{ijb})\frac{\partial E\mu_{ijb}}{\partial \beta_p} = (1-2E\mu_{ijb})\left[\frac{e^{X_{ij}^T\beta_2}X_{ijp}}{(1+e^{X_{ij}^T\beta_2})^2} + \frac{e^{X_{ij}^T\beta_2}X_{ijp}(1-4e^{X_{ij}^T\beta_2}+e^{2X_{ij}^T\beta_2})\sigma_2^2}{4(1+e^{X_{ij}^T\beta_2})^4}\right] \tag{3.26}$$

$$\begin{aligned}\frac{\partial cov(Y_{ijb},Y_{ij'b})}{\partial \beta_p} = {} & \frac{e^{(X_{ij}+X_{ij'})^T\beta_2}\left[X_{ijp}(1+e^{X_{ij'}^T\beta_2})(1-e^{X_{ij}^T\beta_2})+X_{ij'p}(1+e^{X_{ij}^T\beta_2})(1-e^{X_{ij'}^T\beta_2})\right]}{(1+e^{X_{ij}^T\beta_2})^3(1+e^{X_{ij'}^T\beta_2})^3} \\ & \left[\sigma_2^2(\Sigma_{kin})_{ij,ij'} - \frac{(1-e^{X_{ij}^T\beta_2})(1-e^{X_{ij'}^T\beta_2})\sigma_2^4(\Sigma_{kin})_{ij,ij}(\Sigma_{kin})_{ij',ij'}}{4(1+e^{X_{ij}^T\beta_2})(1+e^{X_{ij'}^T\beta_2})}\right] \\ & + \frac{e^{(X_{ij}+X_{ij'})^T\beta_2}\sigma_2^4}{8(1+e^{X_{ij}^T\beta_2})^2(1+e^{X_{ij'}^T\beta_2})^2}\left[\frac{e^{X_{ij}^T\beta_2}X_{ijp}(1-e^{X_{ij'}^T\beta_2})}{(1+e^{X_{ij}^T\beta_2})^2(1+e^{X_{ij'}^T\beta_2})} + \frac{(1-e^{X_{ij}^T\beta_2})e^{X_{ij'}^T\beta_2}X_{ij'p}}{(1+e^{X_{ij}^T\beta_2})(1+e^{X_{ij'}^T\beta_2})^2}\right].\end{aligned} \tag{3.27}$$

A variety of convergence criteria can be applied. For example, one widely used criteria is that the p-norm distance between the parameters of two consecutiveve iterations is smaller than a pre-set threshold. Two example of convergence criteria are given below.

1. $p = 2$, the convergence criteria is Euclidean distance $\sqrt{\sum_{j=1}^{j=p_1+p_2}(\beta_j^{(k)} - \beta_j^{(k+1)})^2 + \sum_{h=1}^{h=m}(\tilde{\alpha}_h^{(k)} - \tilde{\alpha}_h^{(k+1)})^2} <$ pre-specified threshold.
2. $p = \infty$, the criteria becomes $\max(|\beta_1^{(k)} - \beta_1^{(k+1)}|, ..., |\beta_{p_1+p_2}^{(k)} - \beta_{p_1+p_2}^{(k+1)}|, |\tilde{\alpha}_1^{(k)} - \tilde{\alpha}_1^{(k+1)}|, ..., |\tilde{\alpha}_m^{(k)} - \tilde{\alpha}_m^{(k+1)}|) <$pre-specified threshold.

The iterative process continues until convergence.

3.6 Association test

Once parameter estimates have been obtained, it is straightforward to conduct a joint association test between the SNP of interest and two phenotypes. The null hypothesis is $H_0 : \beta_{1SNP} = \beta_{2SNP} = 0$. We perform both Wald test and score test to evaluate the association between the genetic variant and the bivariate phenotypes.

3.6.1 Wald test

The covariance of the parameter estimates is given by

$$cov(\hat{\beta}, \hat{\tilde{\alpha}}) = U^*(\hat{\beta}, \hat{\tilde{\alpha}})^{-1} \sum_{i=1}^{i=N} U_i(\hat{\beta}, \hat{\tilde{\alpha}}) U_i(\hat{\beta}, \hat{\tilde{\alpha}})' U^*(\hat{\beta}, \hat{\tilde{\alpha}})^{-1} \tag{3.28}$$

Hence, a 2-df Wald test can be constructed as follows:

$$\chi^2 = \left(\hat{\beta}_{1SNP}^T, \hat{\beta}_{2SNP}^T\right) var \left(\hat{\beta}_{1SNP}, \hat{\beta}_{2SNP}\right)^{-1} \begin{pmatrix} \hat{\beta}_{1SNP} \\ \hat{\beta}_{2SNP} \end{pmatrix}. \tag{3.29}$$

3.6.2 Score Test

The 2-df score test statistic is formulated as

$$\chi^2 = (\sum_{i=1}^{i=N} U_i^{\beta}(\hat{\beta}_0, \hat{\tilde{\alpha}}_0))^T U^*(\hat{\beta}_0, \hat{\tilde{\alpha}}_0)^{-1}_{\beta_{SNP}\beta_{SNP}} (\sum_{i=1}^{i=N} U_i^{\beta}(\hat{\beta}_0, \hat{\tilde{\alpha}}_0)). \tag{3.30}$$

The subscript 0 indicates that these estimates are obtained under the null hypothesis.

3.7 Simulation studies

We perform simulation studies to evaluate the type-I error under different scenarios, and to compare the Wald test with the score test. We compare power of our approach to other

existing alternatives. In the evaluation of both type-I error rate and power, 1-parameter conditional correlation parametrization is used.

- Type-I error

  Ten thousand datasets are simulated to assess the type-I error rate. In each dataset, FG and BMI with moderate heritability ($h^2 = 0.43$) are simulated based on the covariance matrix estimated from the FHS data. Hence, FG is the quantitative trait and obesity (BMI$>= 30$) is the binary trait in our simulations. The genotypes for 750 nuclear families (2 parents and 2 offspring) are simulated using random allele dropping, independently of the phenotypes.

  Type-I error rate is evaluated for a range of MAF, at a significance threshold of $\alpha = 0.01$. Our simulation results (Table 3.1) demonstrate that the Wald test has elevated

Table 3.1: Type-I error rate evaluated at $\alpha = 0.01$

MAF	Wald (95% CI)	score (95% CI)	MAF	Wald (95% CI)	score (95% CI)
0.3	0.01 (0.08, 0.13)	0.0002 (10^{-5}, 10^{-3})	0.07	0.011 (0.009, 0.014)	0.0066 (0.005, 0.009)
0.2	0.01 (0.007, 0.012)	0.0009 (0.0003, 0.002)	0.05	0.0134 (0.01, 0.017)	0.008 (0.005, 0.01)
0.15	0.0125 (0.0098, 0.016)	0.003 (0.001, 0.005)	0.03	0.0146 (0.012, 0.018)	0.008 (0.006,5 0.01)
0.12	0.011 (0.009, 0.014)	0.003 (0.002, 0.005)	0.01	0.0223 (0.019, 0.026)	0.008 (0.0060, 0.011)
0.1	0.012 (0.0096, 0.015)	0.0052 (0.004, 0.007)	0.005	0.0474 (0.04, 0.05)	0.01 (0.008, 0.014)

  type-I error rate for low-frequency variants (MAF$< 5\%$), while the score test is too conservative for common variants. Therefore, we recommend to use Wald test for common variants (MAF$>= 5\%$) and to use score test for low-frequency variants (MAF$< 5\%$).

- Power calculation

  We simulate another 1000 datasets with 750 nuclear families (2 pareants and 2 offspring) to assess the power of our newly developed tests. Genotypes are simulated using random dropping. The simulation of phenotypes is composed of two steps:

  1. We firstly simulate bivariate continous phenotypes from a multivariate normal

distribution:

$$\begin{pmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\mathbf{1} \\ \mu_\mathbf{2} \end{pmatrix}, \mathbf{\Sigma}\right) \tag{3.31}$$

The mean equations with pre-specified effect sizes are:

$$\begin{cases} \mu_1 = \beta_{01} + 0.05 * age + \beta_1 * g \\ \mu_2 = \beta_{02} + \beta_2 * g \end{cases} \tag{3.32}$$

where

β_{01}, β_{02} are randomly chosen from a uniform distribution $U(-0.1, 0.1)$

$|\beta_1| = \sqrt{\frac{0.5\%}{2q(1-q)}}$, $|\beta_2| = \frac{1}{3}\sqrt{\frac{0.5\%}{2q(1-q)}}$ (q is the sample MAF)

The sign of β_1 and β_2 is randomly determined:

$$P(sgn(\beta_k) = 1) = P(sgn(\beta_k) = -1) = \frac{1}{2} \qquad \forall k = 1, 2 \tag{3.33}$$

with covariance matrix being $\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \otimes \mathbf{\Sigma_{kin}} + \begin{pmatrix} \sigma_{e_1}^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_{e_2}^2 \mathbf{I} \end{pmatrix}$

2. We then convert the second continuous phenotype into a binary phenotype, using a percentile cutoff.

We use the one-parameter conditional correlation matrix in 3.2.2, because it has fewer parameters to estimate and thus takes less time to complete all the simulations. Power is evaluated with respect to two significance thresholds ($\alpha = 0.01, \alpha = 0.001$). We denote p.control the proportion of controls in the simulation of the binary trait (p.control=1-binary trait prevalence). For example, p.control$= 70\%$ means roughly 70% of individuals are controls versus 30% as cases. We investigate the following scenarios:

1. Common Variant (MAF=0.3)

$$p.control = 70\%$$
$$\sigma_1^2 = 1.5; \sigma_2^2 = 2; \sigma_{e_1}^2 = \sigma_{e_2}^2 = 0.5$$
$$|\beta_1| = \sqrt{\frac{0.5\%}{2q(1-q)}} = 3 \times |\beta_2| \tag{3.34}$$

   These variance and proportion parameters are used as default in all other described scenarios, unless specified otherwise.

2. Common Variant (MAF=0.1)
3. Low Frequency Variant (MAF=0.05)
4. Low Frequency Variant (MAF=0.01)
5. MAF=0.3, $\mid \beta_1 \mid = \mid \beta_2 \mid$
6. MAF=0.1, $\mid \beta_1 \mid = \mid \beta_2 \mid$
7. MAF=0.05, $\mid \beta_1 \mid = \mid \beta_2 \mid$
8. MAF=0.01, $\mid \beta_1 \mid = \mid \beta_2 \mid$
9. MAF=0.3, $\mid \beta_2 \mid = 0$
10. MAF=0.1, $\mid \beta_2 \mid = 0$
11. MAF=0.05, $\mid \beta_2 \mid = 0$
12. MAF=0.01, $\mid \beta_2 \mid = 0$.
13. MAF=0.3, $\mid \beta_1 \mid = 0$
14. MAF=0.1, $\mid \beta_1 \mid = 0$

15. MAF=0.05, $\mid \beta_1 \mid = 0$

16. MAF=0.01, $\mid \beta_1 \mid = 0$.

To better compare our method with other commonly used approaches, for each scenario, we compare to the univariate association test adjusted for multiple testing. We select three adjustment methods: Bonferonni, Li and Ji's method [22] and Nyholt's method [31]. The type-I error rate of the minP method is also justified.
We declare statistical significance when the following condition holds:

$$min(p.cont, p.bin) < \alpha_{adj} \tag{3.35}$$

where $\alpha_{adj} = \alpha_{nominal}/N_{eff}$, and N_{eff} is the number of independent tests.

- Bonferonni Correction: $N_{eff} = 2$
- Li and Ji's method:

  $N_{eff} = \sum_{i=1}^{i=M} f(|\lambda_i|)$

  where $f(x) = I(x >= 1) + x - \lfloor x \rfloor$ $(x \geq 0)$
- Nyholt's method:

  $N_{eff} = 1 + (M-1)(1 - \frac{var(\lambda)}{M})$

  where M is the number of tests (i.e. $M = 2$ here). λ_i (λ) are the eigenvalues of the correlation matrix. Although in the original paper, the LD matrix of the loci is used to compute the correlation, here it is adapted to be the correlation matrix for the bivariate phenotypes. Most of the time, Li and Ji's adjustment method yields the same number of effective tests as the Bonferonni adjustment, while Nyholt's method tends to be smaller.

The power results show that at $\alpha = 0.01$ our bivariate approach is at least $\approx 10\%$ more powerful than univariate min p tests when the genetic effect on the two traits

Table 3.2: Power of bivariate tests and univariate tests adjusted for multiple testing

α	MAF	Design	Bivariate	Bonferonni	Li and Ji	Nyholt
0.01	0.01	$\mid \beta_1 \mid = \mid \beta_2 \mid$	0.809	0.75	0.75	0.75
		$\mid \beta_1 \mid = 3 \mid \beta_2 \mid$	0.70	0.68	0.68	0.68
		$\mid \beta_2 \mid = 0$	0.665	0.627	0.627	0.627
		$\mid \beta_1 \mid = 0$	0.371	0.275	0.275	0.275
	0.05	$\mid \beta_1 \mid = \mid \beta_2 \mid$	0.843	0.76	0.76	0.76
		$\mid \beta_1 \mid = 3 \mid \beta_2 \mid$	0.656	0.641	0.641	0.641
		$\mid \beta_2 \mid = 0$	0.637	0.659	0.659	0.659
		$\mid \beta_1 \mid = 0$	0.297	0.351	0.351	0.351
	0.1	$\mid \beta_1 \mid = \mid \beta_2 \mid$	0.811	0.721	0.721	0.721
		$\mid \beta_1 \mid = 3 \mid \beta_2 \mid$	0.614	0.617	0.617	0.617
		$\mid \beta_2 \mid = 0$	0.599	0.626	0.626	0.626
		$\mid \beta_1 \mid = 0$	0.275	0.366	0.366	0.366
	0.3	$\mid \beta_1 \mid = \mid \beta_2 \mid$	0.762	0.681	0.681	0.681
		$\mid \beta_1 \mid = 3 \mid \beta_2 \mid$	0.546	0.554	0.554	0.554
		$\mid \beta_2 \mid = 0$	0.548	0.582	0.582	0.582
		$\mid \beta_1 \mid = 0$	0.24	0.318	0.318	0.318
0.001	0.01	$\mid \beta_1 \mid = \mid \beta_2 \mid$	0.556	0.451	0.451	0.451
		$\mid \beta_1 \mid = 3 \mid \beta_2 \mid$	0.47	0.40	0.40	0.40
		$\mid \beta_2 \mid = 0$	0.423	0.376	0.376	0.376
		$\mid \beta_1 \mid = 0$	0.224	0.106	0.106	0.106
	0.05	$\mid \beta_1 \mid = \mid \beta_2 \mid$	0.623	0.474	0.474	0.474
		$\mid \beta_1 \mid = 3 \mid \beta_2 \mid$	0.382	0.370	0.370	0.370
		$\mid \beta_2 \mid = 0$	0.353	0.391	0.391	0.391
		$\mid \beta_1 \mid = 0$	0.131	0.148	0.148	0.148
	0.1	$\mid \beta_1 \mid = \mid \beta_2 \mid$	0.561	0.436	0.436	0.436
		$\mid \beta_1 \mid = 3 \mid \beta_2 \mid$	0.353	0.348	0.348	0.348
		$\mid \beta_2 \mid = 0$	0.314	0.347	0.347	0.347
		$\mid \beta_1 \mid = 0$	0.116	0.168	0.168	0.168
	0.3	$\mid \beta_1 \mid = \mid \beta_2 \mid$	0.478	0.356	0.356	0.356
		$\mid \beta_1 \mid = 3 \mid \beta_2 \mid$	0.275	0.270	0.270	0.270
		$\mid \beta_2 \mid = 0$	0.285	0.295	0.295	0.295
		$\mid \beta_1 \mid = 0$	0.096	0.156	0.156	0.156

Table 3.3: Power rate when $|\beta_1| = |\beta_2|$ and p.control $= 90\%$ or $= 50\%$

α	MAF	$p.control = 90\%$	$p.control = 50\%$
0.01	0.01	78.20%	86.90%
	0.05	75.40%	85.80%
	0.1	71.50%	82.70%
	0.3	65.70%	77.90%
0.001	0.01	60.80%	70.80%
	0.05	51.90%	64%
	0.1	45.90%	60.60%
	0.3	38.30%	51.30%
0.0001	0.01	44.60%	51.60%
	0.05	31.30%	41.60%
	0.1	26.50%	37.10%
	0.3	19.20%	29.10%

is on the same scale. The power difference becomes less distinguishable when the genetic effect of one trait decreases. Our approach is less powerful in the scenario when there's no dependency on the SNP for the continuous trait. When the significance threshold is increased to $\alpha = 0.001$, we observe similar trends.

We also study the scenarios when the bivariate phenotype has more extreme distribution, such as p.control=90% as well as the scenario in which bivariate phenotype has a balanced distribution, i.e. p.control$= 50\%$. We simulate 1000 datasets and calculate the power of the univariate test under the scenario $|\beta_1| = |\beta_2|$ with all the other parameters set at the default (3.34), and the MAF ranging from 0.01 to 0.3. We observe for each MAF, our approach is the most powerful for the balanced design when the trait prevalence is 50%, and the least powerful when the design is unbalanced (i.e. p.control $= 90\%$ or $= 70\%$). To illustrate the respective power of the quantitative and binary traits, we calculate the power of the univariate tests without adjusting for multiple testing (Table 3.4). For each scenario, the significance of the continuous trait alone accounts for up to 90% of the significance of the bivariate test, while the significance of the binary trait alone accounts for up to 50% of the

significance of the bivariate test.

Table 3.4: Univariate tests in the scenario with equal effect sizes

MAF	α	quantitative	binary	minp[1]	Bivariate
0.01	0.01	73.20%	36.70%	75.00%	80.9%
	0.001	47.20%	13.60%	45.10%	55.6%
0.05	0.01	71.70%	43.00%	76.00%	84.3%
	0.001	44.60%	22.10%	47.40%	62.3%
0.1	0.01	69.70%	39.40%	72.10%	81.1%
	0.001	41.60%	17.70%	43.60%	56.1%
0.3	0.01	63.00%	37.50%	68.10%	76.2%
	0.001	34.20%	16.60%	35.60%	47.8%

3.8 Binary trait association test based on EGEE

Our approach can also be restricted to study the association between the genetic variant and the binary trait. We conduct some simulation studies to evaluate the type-I error of our approach when restricting to binary trait alone in family samples in the following scenarios (Table 3.5).

3.9 Data analysis

We apply our approach to study the association between the genetic variants on chromosome 16 and the bivariate phenotypes of BMI and T2D status, because some genes on chromosome 16 such as *FTO* are known to be strongly associated with BMI and T2D [19] [18] [6] [47] [55] [56].

3.9.1 Phenotype dataset

We use the FHS dataset. The FHS was initiated in 1948 and is a longitudinal study consisting of three generations of cohorts: the Original cohort, the Offspring cohort and the 3rd generation (Gen 3) cohort, comprising up to 14428 participants, some recruited from

[1]In each simulation, the smallest effective number of phenotypes is always chosen among the three multiple adjustment methods.

Table 3.5: Type-I error rate evaluated at $\alpha = 0.01$

MAF	typeIerror	lower 95% CI	upper 95% CI
0.3	0.0104	0.0085	0.0126
0.25	0.0116	0.0096	0.0139
0.2	0.0118	0.0098	0.014
0.15	0.0102	0.0083	0.012
0.1	0.0112	0.009	0.0135
0.05	0.0106	0.0087	0.013
0.045	0.0086	0.0069	0.011
0.04	0.0075	0.0059	0.0094
0.035	0.0077	0.0061	0.01
0.03	0.008	0.0063	0.01
0.025	0.0077	0.0061	0.01
0.02	0.0079	0.0063	0.01
0.015	0.0109	0.009	0.013
0.01	0.0077	0.0061	0.01
0.005	0.0099	0.008	0.012

the same family and hence related. All the participants are from the town of Framingham, Massachusetts. Over the years, FHS have been successful in identifying risk factors of cardiovascular-related traits like blood pressure, cholesterol level as well as risk factors for glycemic and metabolic traits. We select BMI and T2D status at exam 7 as the bivariate phenotypes. The association tests with BMI are adjusted for age, but T2D analyses are not adjusted for age. There is a total of 8384 genotyped and phenotyped participants in the analysis. After excluding participants with missing phenotype or covariates and large families (famsize>50), we end up with 5676 individuals in 1172 families.

Table 3.6: The characteristics of the phenotype dataset

gender	N	average BMI	proportion of T2D cases	average age (years)
male	2679	27.98	6.34%	46.92
female	2997	26.26	3.77%	46.67

3.9.2 Genotype

We use the imputed genotype of SNP Health Association Resource (SHARe) and we analyze chromosome 16 only. The NHLBI SHARe Project includes GWAS for several large, multi-ethnic NHLBI Cohort studies of men and women to identify genes underlying cardiovascular and lung disease and other disorders like osteoporosis and diabetes. The Framingham SHARe was the first cohort released in Oct 2007, with genotypes from over 550,000 SNPs in over 9,200 participants. Additional SNPs were imputed with the software MACH (Markov Chain based haplotyper) developed by researchers from the University of Michigan using the HapMap 2 reference haplotypes.

3.9.3 Correlation

The one-parameter conditional correlation matrix (section 3.2.2) is used for the sake of less computational time.

3.9.4 Results

We list the top 20 SNPs on chromosome 16 in Table 3.6, out of which 15 variants are in the *FTO* gene known to be associated with both BMI and T2D status [19] [18] [6] [47] [55] [56]. To adjust for multiple testing, the number of independent SNPs is calculated using Li and Ji's method [22]. Thus, the adjusted significance threshold equals $0.05/5900 = 8.5 \times 10^{-6}$ and all the top 20 SNPs fall below the threshold. The top 3 SNPs are on *ADCY9* gene which is also known to be associated with BMI [40]. Moreover, one variant (rs2303220) from the top 20 is on *PKD1L3*, a gene associated with a kidney disease. This is an interesting finding because people with T2D have a higher chance of developing some kidney disease eventually. People with a kidney disease are more likely to suffer from some metabolic disorder, thus tend to have lower BMI level.

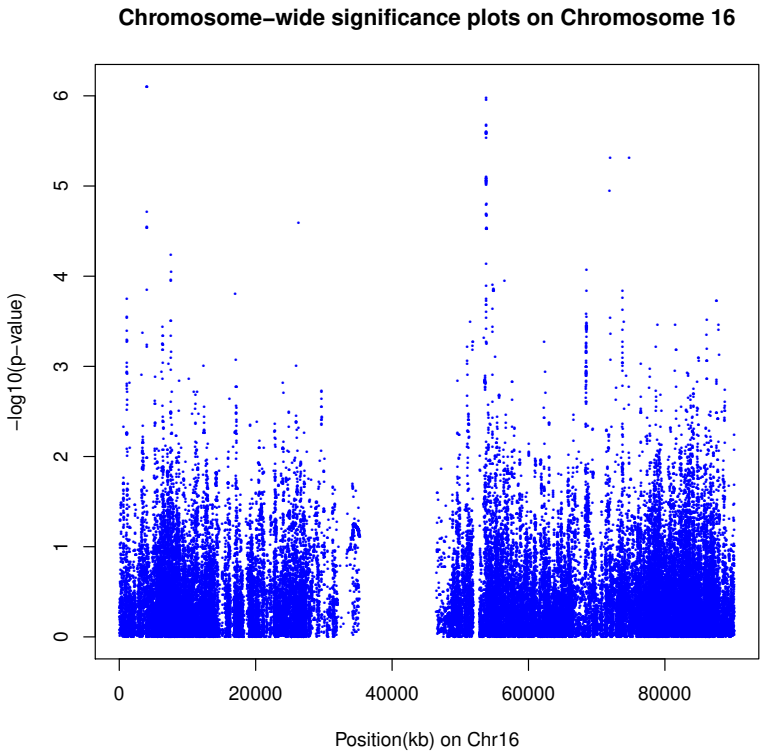

Figure 3-1: Chromosome-wide significance of SNPs on Chromosome 16 and their association with BMI and T2D status

3.10 Discussion

We propose a novel approach to test the association between a genetic variant and binary phenotypes in family samples, based on EGEE. Our approach can handle a range of pedigrees including large and complex pedigrees. From extensive simulation studies, we demonstrate our approach has the correct type-I error rate in the scenarios evaluated, and is consistently more powerful than univariate tests adjusted for multiple testing in certain scenarios.

Our approach is based on quasi-likelihood. Fisher's scoring algorithm is implemented for parameter estimation, thanks to its advantage of fast convergence. As an important

Table 3.7: Top 20 results for Chromosome 16

dbSNP	Position	Iter	p_value	Type
rs12448453	4065412	8	7.9E-07	Wald
rs2239311	4066191	8	7.9E-07	Wald
rs2238452	4068563	8	8.0E-07	Wald
rs1558902	53803574	9	1.1E-06	Wald
rs1421085	53800954	9	1.1E-06	Wald
rs9940646	53800629	9	2.1E-06	Wald
rs11075985	53805207	9	2.1E-06	Wald
rs9923544	53801985	9	2.5E-06	Wald
rs9923147	53801549	9	2.5E-06	Wald
rs9930333	53799977	9	2.6E-06	Wald
rs9940128	53800754	9	2.6E-06	Wald
rs9939973	53800568	9	2.6E-06	Wald
rs9928094	53799905	9	2.7E-06	Wald
rs9937053	53799507	9	2.7E-06	Wald
rs1121980	53809247	9	2.9E-06	Wald
rs7198396	74793644	NA	4.8E-06	score
rs2303220	71988728	9	4.8E-06	Wald
rs9936385	53819169	9	8.0E-06	Wald
rs7193144	53810686	9	8.1E-06	Wald
rs9939609	53820527	9	8.3E-06	Wald

component of the Information matrix, the covariance matrix of the bivariate phenotypes is derived using second-order taylor expansion approximation, with respect to the random effects accounting for the familial correlation. Despite its complex form, the higher-order taylor expansion is more precise than the delta's method. It might be worth exploring in the future the added value of using an expansion with order higher than two.

We propose to use a conditional correlation matrix to account for the correlation of the continuous and the binary traits for any pair of individuals. The two types of conditional correlation matrices do not differ much in terms of the computational efficiency. However, it remains to be evaluated if the difference between these two parameterizations is negligible in terms of parameter estimation and type-I error rate.

Using Fisher's scoring algorithm assuming no over or under dispersion, we estimate correlation parameters as well as regression parameters (effect size) simultaneously. Although in the current model we assume no over- or under- dispersion for the binary trait, it can be easily incorporated if we want to take into consideration the possibility that over- or under- dispersion occurs.

We perform both Wald and score tests in the simulation studies to evaluate the type-I error rate with respect to all the different MAF scenarios. We conclude Wald test has the correct type-I error rate for common variants but elevated type-I error rate for low-frequency variants (MAF$< 5\%$). On the contrary, we find that the score test gives the correct type-I error rate for low-frequency variants (MAF$< 5\%$), but seems too conservative for common variants.

We compare the power of our approach to the min p method of univariate tests adjusted for multiple testing, in a number of different scenarios with MAF ranging from 0.01 to 0.3. Our simulation results show our approach is consistently more powerful, with the power maximized in the scenario where the SNP effect on both traits are equivalently strong.

Chapter 4

Haplotype association analysis and meta-analysis

4.1 Background

GWAS have identified 56 common and mostly non-exonic SNPs associated with FG and FI levels. They together explain 4.8% of the FG variation and 1.2% of the variance in FI [39]. Our recent large-scale exome-chip meta-analysis study, comprising up to 24 CHARGE cohorts with European ($N_{max} = 50900$) and African ($N_{max} = 9664$) samples identified association between FG and rare variants in G6PC2, a known FG-associated locus with a common variant, rs560887, discovered in prior GWAS. To further understand the association between the genetic architecture of a region and a trait, we develop a meta-analysis approach to evaluate the association between haplotypes formed by multiple SNPs in a region and FG. Meta-analysis has been used by large consortia to improve power by increasing sample size. For example, meta-analysis has been widely used in single-variant or gene-based tests. However, for haplotype analysis, there are no available methods due to some challenges: the haplotypes observed by different cohorts or ethnic groups might be different; the haplotype structure can become more complex, with an increasing number of variants in a region. We propose a two-stage approach, to address this question efficiently. In the first stage, each cohort computes the expected haplotype effects

in a regression framework including a random familial effect to account for the relatedness, if appropriate. For the second stage, we propose a multivariate generalized least square meta-analysis approach to combine haplotype effects from multiple cohorts. Association tests for each haplotype and a global test can be obtained within our framework. Simulation studies show our approach achieves the correct type-I error rate.

4.2 Methods

4.2.1 Single cohort haplotype association test

Our approach is based on a method developed by Zaykin et al. [53] for unrelated samples. However, we incorporates family structure, making it applicable to both unrelated and related samples. The general model for K observed haplotypes is

$$Y = X\gamma + \beta_1 h_1 + ... + \beta_K h_K + b + \epsilon \tag{4.1}$$

Where

Y is a quantitative trait;

X is the covariates matrix (without intercept);

h_m $(m = 1, ..., K)$ is the expected haplotype dosage: when the haplotype is observed, the value is 0 or 1or 2; otherwise, expected haplotype dosage is statistically inferred from genotype. So for each row (subject), the summation of h_m $(m = 1, ..., K)$ is always 2;

b is the random effect to account for the family structure (if related individuals are present in the sample), and is set to 0 for unrelated samples;

ϵ is the random error term.

4.2.2 Meta-Anlaysis

Haplotype association results are obtained for each cohort. We use meta-analysis to combine the information from each cohort to maximize the sample size. Multivariate

meta-analysis is applied to summarize association findings. The vector of parameters for the haplotype effect is modeled as follows:

$$\beta = \begin{pmatrix} \beta^1 \\ \vdots \\ \beta^N \end{pmatrix} = W\beta_{meta} + e = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta^1_{meta} \\ \vdots \\ \beta^{K_{total}}_{meta} \end{pmatrix} + e \tag{4.2}$$

where

β^i ($i = 1, \cdots, N$) is the haplotype coefficient vector of cohort i;

W is the design matrix specifying which haplotypes each cohort contributes out of K_{total} distinct haplotypes;

β_{meta} is the coefficient vector of the meta-analysis;

K_{total} is the total number of distinct haplotypes contributed at least one cohort;

e is the error term which has a multivariate normal distribution with a mean of 0 and a diagonal covariance matrix of $\Sigma = \begin{pmatrix} var(\beta^1) & \cdots & 0 \\ \vdots & var(\beta^k) & \vdots \\ 0 & \cdots & var(\beta^N) \end{pmatrix}$.

One of the advantages is to allow each cohort to contribute unique haplotypes in addition to haplotypes that are observed in multiple cohorts. The best linear unbiased estimator (BLUE) of β_{meta} is $\hat{\beta_{meta}} = (W^T\Sigma^{-1}W)^{-1}W^T\Sigma^{-1}\beta$, and the variance of $\hat{\beta_{meta}}$ is $Var(\hat{\beta_{meta}}) = (W^T\Sigma^{-1}W)^{-1}$. Because the covariance matrix Σ is always unknown, we substitute the sample estimate $\hat{\Sigma}$, then $\hat{\beta_{meta}} = (W^T\hat{\Sigma}^{-1}W)^{-1}W^T\hat{\Sigma}^{-1}\beta$, and $U = Var(\hat{\beta_{meta}}) = (W^T\hat{\Sigma}^{-1}W)^{-1}$.

4.2.3 Hypothesis testing

The null hypothesis of no haplotype association is expressed as

$$H_0 : \beta^1_{meta} = \beta^2_{meta} = ... = \beta^{K_{total}}_{meta}. \tag{4.3}$$

To construct a test statistic, we first reparameterize the haplotype effect parameters, and the equivalent null hypothesis becomes:

$$H_0 : \begin{pmatrix} \gamma^2_{meta} \\ \vdots \\ \gamma^{K_{total}}_{meta} \end{pmatrix} = \begin{pmatrix} \beta^2_{meta} - \beta^1_{meta} \\ \vdots \\ \beta^{K_{total}}_{meta} - \beta^1_{meta} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.4}$$

The null hypothesis can be tested using a Wald test statistic of the form

$$\chi^2 = \hat{\gamma_{meta}}^T V^{-1}(\hat{\gamma_{meta}})\hat{\gamma_{meta}} \tag{4.5}$$

where the elements of V is expressed as $V_{jj'} = U_{jj'} - U_{j1} - U_{1j'} + U_{11}$. The Wald test statistic follows a $\chi^2_{Ktotal-1}$ asymptotically.

4.3 Type-I error and power evaluation

4.3.1 Type-I error

Ten thousand simulations are performed to assess the type-I error rate. We simulate a trait with moderate heritability ($h^2 = 20\%$) and with the following distribution $Y \sim N(\mu, \Sigma)$

where $\Sigma = \sigma^2_a \Phi + \sigma^2_e I = 2\sigma^2_a \Sigma_{kin} + \sigma^2_e I$ with $\sigma^2_a = 0.5$

We choose the known T2D associated gene *G6PC2* (chromosome 2) to generate the reference panel of haplotype frequencies using FHS exome-chip data. We simulate

the haplotypes of founders based on the estimated haplotype frequency panel of FHS and obtain the genotypes of the offspring by random haplotype dropping assuming no recombination occurs in haplotypes. The type-I error rate is evaluated in the scenarios listed in (Table 4.2).

Table 4.1: G6PC2 variants

Probe Name	Location	rsID	REF	ALT	MAF
exm239638	169757930	rs142189264	T	C	0.00066
exm239639	169757953	rs149874491	C	A	0.00021
exm239642	169758029	rs201561079	C	T	0.000025
exm239643	169758044	rs199682245	T	A	0.00013
exm239662	169763244	rs2232322	G	A	0.00027
exm239663	169763245	rs145050507	C	T	0.00047
exm239664	169763262	rs138726309	T	C	0.0036
exm239667	169764141	rs2232323	C	A	0.0078
exm239672	169764176	rs492594	C	G	0.46
exm239675	169764210	rs145217135	C	T	0.000037
exm239684	169764287	rs150538801	C	T	0.00061
exm239690	169764368	rs146779637	T	C	0.0028
exm239695	169764449	rs200336133	T	C	0.000025
exm239698	169764491	rs2232326	C	T	0.0018
exm-rs560887	169763148	rs560887	A	G	0.29

Table 4.2: Scenarios for Type-I error evaluation

Scenario	*N_cohort	SampleSize_each_cohort
1	5	250 families (*famsize=4)×5
2	5	1000 unrelated ×5
3	5	400, 700, 1000, 1300, 1600 unrelated
4	5	200, 200, 200, 200, 450 families (famsize=4)
5	5	100, 175, 250, 325, 400 families (famsize=4)
6	10	250 families (famsize=4)×5; 1000 unrelated ×5
7	10	250, 125, 125, 375, 375 families (famsize=4); 1000, 500, 500, 1500, 1500 unrelated
8	5	250 families (famsize in c(3,4,5,6))
9	5	100, 175, 250, 325, 400 families (famsize in c(3,4,5,6))
10	10	250 families (famsize in c(3,4,5,6))×5; 5 unrelated cohorts has the same samples size with the 5 family cohorts
11	10	250 families (famsize in c(3,4,5,6))×7; 1000 unrelated ×3

* N_cohort: the number of cohorts simulated;

* famsize: the number of subjects in each family

Table 4.3: Type-I error results

α	Scenario	Type-I error Rate	upper 95%CI	lower 95%CI
0.01	1	0.010	0.0085	0.013
	2	0.010	0.0085	0.013
	3	0.0095	0.0077	0.012
	4	0.011	0.009	0.013
	5	0.011	0.009	0.013
	6	0.01	0.008	0.012
	8	0.0094	0.008	0.011
	9	0.011	0.009	0.013
	10	0.0105	0.009	0.013
	11	0.009	0.007	0.011

4.3.2 Power calculation

Ten thousand simulations with 5 or 10 (depending on the scenario) independent cohorts are simulated to assess the power of our approach. We first select a known T2D associated gene *JAZF1* (chromosome 7) and generate the reference panel of haplotype frequencies from FHS exome-chip data. There is no single haplotype dominating the structure of *JAZF1* (Table 4.5). Instead, at least 8 haplotypes are required to represent the genetic structure of this region. There are 20 haplotypes observed in FHS data, because all the 5 variants (Table 4.4) are common variants, when simulating the genotype of nuclear families using random dropping there are 32 haplotypes observed in total. As a result, it is not possible to use a few haplotypes to represent the haplotype structure in this region. We simulate the haplotypes of the founders and then generate the genotypes of the offspring using random haplotype dropping assuming no recombination. An age variable is simulated to be used as a covariate: for each family, we first simulate the age of the founders from a discrete uniform distribution of U(25, 90), and then generate the age of the offspring via a discrete uniform distribution of U(1, min(founder_age)-20).

Power is evaluated in the following four scenarios (phenotype datasets), with varying haplotype and SNP effects. The four phenotype datasets share the same covariance matrix

Table 4.4: JAZF1 variants

Probe	Chr	MapInfo	rsID	SKATgene	Minor	Major	MAF
exm-rs10486567	7	27976563	rs10486567	JAZF1	A	G	0.2415
exm2270592	7	28039797	rs38523	JAZF1	C	T	0.3683
exm-rs864745	7	28180556	rs864745	JAZF1	G	A	0.4965
exm-rs1635852	7	28189411	rs1635852	JAZF1	C	T	0.4973
exm-rs849134	7	28196222	rs849134	JAZF1	G	A	0.4917

Table 4.5: JAZF1 haplotype frequencies

Haplotype	rs10486567	rs38523	rs864745	rs1635852	rs849134	Haplotype Frequency
hap1	G	T	A	T	A	0.232702
hap2	G	T	G	C	G	0.229488
hap3	G	C	G	C	G	0.160802
hap4	G	C	A	T	A	0.129456
hap5	A	T	A	T	A	0.086615
hap6	A	T	G	C	G	0.079319
hap7	A	C	A	T	A	0.043359
hap8	A	C	G	C	G	0.025893
hap9	A	T	G	T	A	0.002862
hap10	A	T	A	C	A	0.002855
hap11	A	C	A	C	A	0.00231
hap12	G	T	A	C	A	0.001862
hap13	G	T	G	T	A	0.001658
hap14	G	C	G	T	A	0.000526
hap15	A	C	G	T	G	< 0.0001
hap16	A	C	A	C	G	< 0.0001
hap17	A	T	A	T	G	< 0.0001
hap18	G	T	G	C	A	< 0.0001
hap19	G	C	A	C	A	< 0.0001
hap20	A	C	G	T	A	< 0.0001

defined as $\mathbf{Var}(\mathbf{Y}) = \sigma_a^2\mathbf{\Phi} + \sigma_e^2\mathbf{I}$

where $\sigma_e^2 = 0.5$, $\sigma_a^2 = 0.5$, Φ is twice the kinship matrix and hence heritability h^2 is set to 0.5

We select two haplotypes, GTATA (the most frequent haplotype) and GCGCG (the third most frequent haplotype), to have an effect on the phenotype while all other haplotypes have no effect on the phenotype. For models with SNP effect only, we select rs849134 and rs38523 to have non-zero effect on the trait while all other SNPs have no effect on the

trait. The conditional mean phenotype (conditional on the observed haplotypes or SNPs) is based on the following equations, where the phenotype is influenced by 1 haplotype, 2 haplotypes, 1 SNP or 2 SNPs depending on the scenarios:

1. One haplotype effect: $\hat{\mu} = \mathbf{age} \times 0.05 + \mathbf{h_1}\sqrt{\frac{R^2}{\bar{h_1}(1-\frac{\bar{h_1}}{2})}}$
   hap1(the most frequent) is selected as the predictor.
2. Two haplotype effects: $\hat{\mu} = \mathbf{age} \times 0.05 + \mathbf{h_1}\sqrt{\frac{R^2}{2\bar{h_1}(1-\frac{\bar{h_1}}{2})}} + \mathbf{h_3}\sqrt{\frac{R^2}{2\bar{h_3}(1-\frac{\bar{h_3}}{2})}}$
   hap1 and 3 are selected as the predictors.
3. One SNP effect: $\hat{\mu} = \mathbf{age} \times 0.05 + \mathbf{SNP_5}\sqrt{\frac{R^2}{2MAF_5(1-MAF_5)}}$
   rs849134 is selected as the predictor.
4. Two SNP effects: $\hat{\mu} = \mathbf{age} \times 0.05 + \mathbf{SNP_5}\sqrt{\frac{R^2}{4MAF_5(1-MAF_5)}} + \mathbf{SNP_2}\sqrt{\frac{R^2}{4MAF_2(1-MAF_2)}}$
   rs38523 and rs849134 are selected as the predictors.

Note that h_1 and h_3 are the dosage of the most and the 3rd most frequent haplotypes;
$\bar{h_1}$ and $\bar{h_3}$ are the average dosage of the most and the 3rd most frequent haplotypes, based on all the samples;
R^2 is the proportion of variance explained by the haplotypes: we choose $R^2 = 1\%$;
SNP_5, SNP_2, are the SNP dosage of rs849134 and rs38523 using additive coding;
and MAF_5, MAF_2, are the sample MAF of rs849134 and rs385235.
The trait is then simulated from a multivariate normal distribution with conditional mean and covariance matrix specified as above, i.e., $\mathbf{Y} \sim \mathbf{N}(\hat{\mu}, \mathbf{Var}(\mathbf{Y}))$.

For each haplotype (SNP) scenario, we generate four cohort scenarios, with fixed

Table 4.6: Cohort scenario for power assessment

Cohort Scenario	N_cohort	SampleSize_each_cohort
I	5	250 families (famsize=4)×5
II	5	250 families (famsize in c(3,4,5,6))
III	5	100, 175 families (famsize=4); 400, 700, 1000 unrelated
IV	5	100, 175 families (famsize in c(3,4,5,6)); 400, 700, 1000 unrelated

and varying family size, all family-based cohorts and a mixture of family-based and unrelated cohorts (Table 4.6). So it ends up with a total of 16 scenarios. In each scenario, we evaluate the power of our approach and compare with single SNP tests adjusted for multiple testing. For our approach, we first implement the haplotype association test for each cohort and then summarize the results through meta-analysis. For the SNP effect, we perform meta-analysis of single-SNP tests, calculate the minimum p-value (min P), and adjust it for multiple testing using Gao's methods [13] recommended by Hendricks et al. [17]. The significance threshold $\alpha = 0.01$ is used.

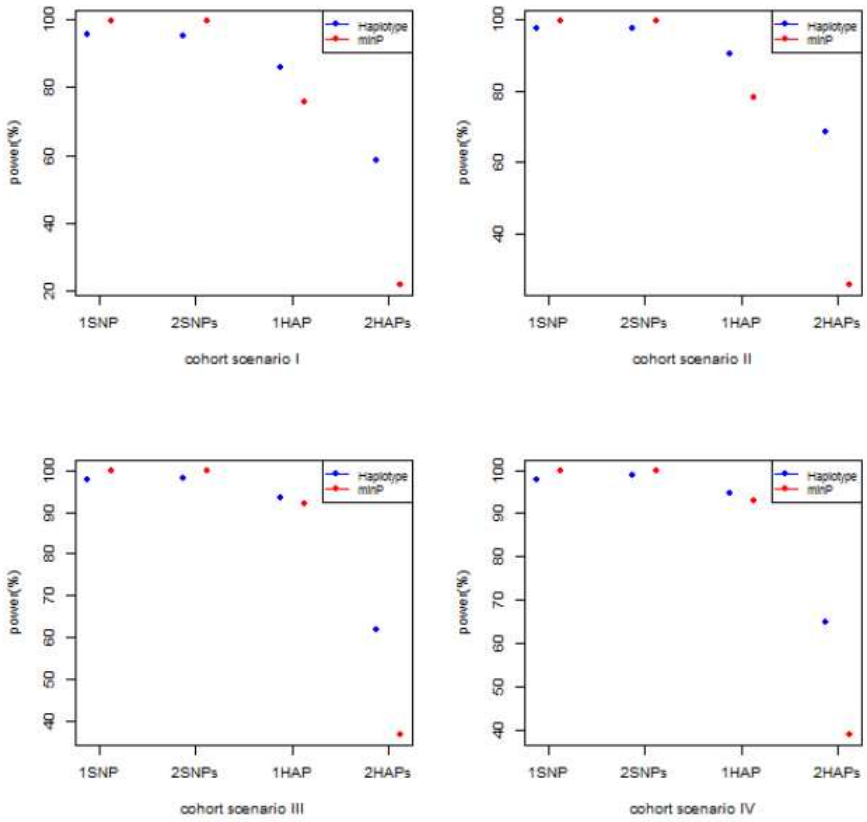

Figure 4-1: Power of the haplotype meta-analyis approach compared to single SNP meta-analysis using the minimum P-value adjusted for multiple testing evaluated at $\alpha = 0.01$, with respect to the 4 cohorts scenarios

Across the four cohort scenarios, we observe similar contrasting pattern between our method and the min P method. In the phenotype scenario where the phenotypes are simulated directly from SNPs, we observe the two methods have approximately the same power, although intuitively the min P method is thought to be more powerful, due to the stronger connection between the SNPs and the phenotypes. Furthermore, our approach is obviously more powerful in the phenotypes simulated from haplotypes, especially for the phenotypes simulated from two haplotypes. For example, our approach is $\sim 25\% - 40\%$ more powerful than the min P method when the phenotypes are simulated from two haplotypes.

4.4 Data analysis

Our approach is applied to study the association between the *G6PC2* haplotype structure and FG, based on the CHARGE exome-chip data. There are 17 exonic variants in the G6PC2 region, 15 rare variants (MAF$< 1\%$) and 2 common variants (rs560887 with MAF=25.4%; rs492594 with MAF=43.7%). Previous GWAS have identified rs560887, one of the two common variants, to be associated with FG level. A recent large-scale exome-chip analysis further indicated that the joint analysis of these 15 rare variants show association with FG. To understand how the haplotype structure of these 15 rare variants alone or with rs560887 impact FG level, we perform 2 sets of meta-analysis, after collecting haplotype association results from 18 European CHARGE cohorts, comprising up to 38322 non-diabetic participants.

4.4.1 Single cohort haplotype association test

Preliminary analyses indicates that the most common haplotype is shared by all 18 cohorts, so we select the most frequent haplotype as the reference haplotype and specifically formulate the model without the term of the reference haplotype:

$Y = \mu + X\gamma + \beta_2 h_2 + ... + \beta_K h_K + b + \epsilon$

where

μ is the intercept;

Y is the trait;

X is the covariates matrix;

h_m $(m = 2, ..., K)$ is the expected haplotype dosage (if the haplotype is observed, the value is 0 or 1 or 2; otherwise, it is statistically inferred from genotype);

b is the random effect to account for the family structure (if present), and is set to 0 for unrelated samples;

ϵ is the random error.

4.4.2 Meta-Analysis

We apply the same meta-analysis approach as in 4.2.2 to the estimates at cohort level.

4.4.3 Hypothesis testing

The null hypothesis is $H_0 : \beta^2_{meta} = ... = \beta^{K_{total}}_{meta} = 0$.

The Wald-Test statistic is constructed without additional reparametrization as $\chi^2 = \hat{\beta_{meta}}^T V^{-1}(\hat{\beta_{meta}})\hat{\beta_{meta}}$ and follows a $\chi^2_{K_{total}-1}$ asymptotically.

Table 4.7: G6PC2 variants

Name	Chr	MapInfo	dbSNPID	Minor	Major	MAF
exm239638	2	169757930	rs142189264	T	C	0.00041
exm239639	2	169757953	rs149874491	C	A	0.00017
exm239642	2	169758029	rs201561079	C	T	0.00011
exm239643	2	169758044	rs199682245	T	A	0.00007
exm239650	2	169761057	rs187707963	G	A	0.00006
exm-rs560887	2	169763148	rs560887	A	G	0.25424
exm239662	2	169763244	rs2232322	G	A	0.00017
exm239663	2	169763245	rs145050507	C	T	0.00074
exm239664	2	169763262	rs138726309	T	C	0.00318
exm239667	2	169764141	rs2232323	C	A	0.00611
exm239672	2	169764176	rs492594	C	G	0.43656
exm239675	2	169764210	rs145217135	C	T	0.00004
exm239682	2	169764269	rs147360987	T	C	0.0001
exm239684	2	169764287	rs150538801	C	T	0.00044
exm239687	2	169764338	rs148689354	G	A	0.00032
exm239690	2	169764368	rs146779637	T	C	0.00253
exm239698	2	169764491	rs2232326	C	T	0.0019

4.4.4 Results

Two meta-analyses are performed: one with the rare variants only, and one with the rare variants plus the GWAS-identified common variant rs560887.

Rare Variants Only

The global haplotype association test has a p-value of 1.1×10^{-17}, with the most frequent haplotype having an overall frequency of 98.2%, which is more significant than

any single SNP association test, SKAT and burden test [49]. In other words, it reinforces the conclusions of exome-chip results that the 15 rare variants of G6PC2 region play a vital role in influencing FG level. We also test for individual haplotype effects: the most significant haplotypes are the ones carrying a single rare allele at the exm239690 variant ($p = 2.84 \times 10^{-10}$), the exm239698 variant ($p = 1.4 \times 10^{-7}$) and the exm239667 variant ($p = 1.45 \times 10^{-6}$). This is consistent with the findings of the single-variant tests.

Table 4.8: Haplotype analysis of G6PC2 rare variants

exm239638	exm239639	exm239642	exm239643	exm239650	exm239662	exm239663	exm239664	exm239667	exm239675	exm239682	exm239684	exm239687	exm239690	exm239698	hapcat	freq	beta_meta	p.ind
C	A	T	A	A	A	T	C	A	T	C	T	A	C	T	1	18	NA	NA
C	A	T	A	A	A	T	C	C	T	C	T	A	C	T	2	18	-0.11	1.5×10^{-6}
C	A	T	A	A	A	T	C	A	T	C	T	A	T	T	3	17	-0.22	2.8×10^{-10}
C	A	T	A	A	A	T	T	A	T	C	T	A	C	T	4	16	-0.09	0.02
C	A	T	A	A	A	T	C	A	T	C	T	A	C	C	5	13	-0.26	1.4×10^{-7}
C	A	T	A	A	A	T	C	A	T	C	C	A	C	T	6	11	-0.13	0.22
C	A	T	A	A	A	C	C	A	T	C	T	A	C	T	7	11	-0.07	0.44
T	A	T	A	A	A	T	C	A	T	C	T	A	C	T	8	10	-0.22	0.03
C	A	T	A	A	G	T	C	A	T	C	T	A	C	T	9	7	0.22	0.13
C	C	T	A	A	A	T	C	A	T	C	T	A	C	T	10	3	-0.19	0.14
C	A	T	A	A	A	T	T	C	T	C	T	A	C	T	11	3	-0.89	0.00
C	A	T	T	A	A	T	C	A	T	C	T	A	C	C	12	3	-0.21	0.70
C	A	C	A	A	A	T	C	A	T	C	T	A	C	T	13	2	0.57	0.22
T	A	T	A	A	A	T	C	A	T	C	T	A	C	C	14	1	0.21	0.64
C	A	T	A	G	A	T	C	A	T	C	T	A	C	T	15	1	-0.48	0.41
C	A	T	A	NA	A	T	C	A	C	C	T	A	C	T	16	1	0.91	0.42
C	A	T	NA	A	A	T	C	A	C	C	T	A	C	T	17	1	0.10	0.83
C	A	T	NA	A	A	T	T	A	T	C	T	A	T	T	18	1	1.31	0.01
C	A	T	A	A	A	T	T	A	T	C	T	A	C	C	19	1	-0.73	0.59
C	A	T	A	A	A	T	C	C	T	C	T	A	C	C	20	1	-1.10	0.44
C	A	T	T	A	A	T	C	A	T	C	T	A	C	T	21	1	-0.52	0.14

Rare Variants+rs560887

Adding the common variant rs560887 to the haplotype analysis results in a more significant global haplotype association test (p=1.5×10^{-81}), with the most significant haplotype carrying the minor allele C at the common variant rs560887 (Table 4.9).

Table 4.9: Haplotype analysis of G6PC2 rare variants

exm239638	exm239639	exm239642	exm239643	exm239650	exm-rs560887	exm239662	exm239663	exm239664	exm239667	exm239675	exm239682	exm239684	exm239687	exm239690	exm239698	hapcat	freq	beta_meta	p.ind
C	A	T	A	A	C	A	T	C	A	T	C	T	A	C	T	1	18	NA	NA
C	A	T	A	A	T	A	T	C	A	T	C	T	A	C	T	2	18	-0.08	8.9×10^{-77}
C	A	T	A	A	T	A	T	C	C	T	C	T	A	C	T	3	18	-0.13	7.7×10^{-9}
C	A	T	A	A	T	A	T	C	A	T	C	T	A	T	T	4	17	-0.24	5.6×10^{-12}
C	A	T	A	A	C	A	T	T	A	T	C	T	A	C	T	5	16	-0.11	7.0×10^{-3}
C	A	T	A	A	C	A	T	C	A	T	C	T	A	C	C	6	13	-0.28	2.0×10^{-8}
C	A	T	A	A	C	A	T	C	A	T	C	C	A	C	T	7	11	-0.19	0.10
T	A	T	A	A	C	A	T	C	A	T	C	T	A	C	T	8	10	-0.28	0.01
C	A	T	A	A	C	A	C	C	A	T	C	T	A	C	T	9	9	-0.08	0.43
C	A	T	A	A	T	A	T	T	A	T	C	T	A	C	T	10	9	-2.51	0.30
C	A	T	A	A	T	A	T	C	A	T	C	T	A	C	C	11	8	-0.22	0.66
C	A	T	A	A	T	A	C	C	A	T	C	T	A	C	T	12	7	-0.38	0.40
T	A	T	A	A	T	A	T	C	A	T	C	T	A	C	T	13	6	0.11	0.81
C	A	T	A	A	C	G	T	C	A	T	C	T	A	C	T	14	6	0.32	0.17
C	A	T	A	A	T	G	T	C	A	T	C	T	A	C	T	15	5	0.06	0.79
C	A	T	A	A	C	A	T	C	C	T	C	T	A	C	T	16	4	-0.32	0.58
C	A	T	A	A	T	A	T	C	A	T	C	C	A	C	T	17	4	0.13	0.74
C	C	T	A	A	C	A	T	C	A	T	C	T	A	C	T	18	3	-0.18	0.23
C	A	T	A	A	T	A	T	T	C	T	C	T	A	C	T	19	3	-0.95	0.00
C	A	T	T	A	C	A	T	C	A	T	C	T	A	C	C	20	3	-0.25	0.64
C	A	T	A	A	C	A	T	C	A	T	C	T	A	T	T	21	2	-619118.00	0.76
C	A	C	A	A	C	A	T	C	A	T	C	T	A	C	T	22	2	0.51	0.26
T	A	T	A	A	C	A	T	C	A	T	C	T	A	C	C	23	1	0.16	0.71
C	C	T	A	A	T	A	T	C	A	T	C	T	A	C	T	24	1	-199.97	0.57
C	A	T	A	G	C	A	T	C	A	T	C	T	A	C	T	25	1	-0.53	0.36
C	A	T	A	NA	C	A	T	C	A	C	C	T	A	C	T	26	1	0.86	0.44
C	A	T	NA	A	C	A	T	C	A	C	C	T	A	C	T	27	1	0.05	0.91
C	A	T	NA	A	T	A	T	T	A	T	C	T	A	T	T	28	1	1.29	0.01
C	A	T	A	A	C	A	T	T	A	T	C	T	A	C	C	29	1	-0.78	0.57
C	A	T	A	A	T	A	T	C	C	T	C	T	A	C	C	30	1	-1.13	0.43
C	A	T	T	A	C	A	T	C	A	T	C	T	A	C	T	31	1	-0.45	0.37
C	A	T	T	A	T	A	T	C	A	T	C	T	A	C	T	32	1	-0.96	0.53

4.5 Discussion

Here we propose a general meta-analysis approach to combine the results of haplotype association test from cohorts. Our approach has no restrictions on the haplotypes observed across the cohorts. Instead, we allow cohorts to contribute unique haplotypes in addition to those observed in common. In the first stage, cohorts can use our existing scripts to perform association test at the cohort level, and then send back the effect size estimate and covariance matrix to the central analyst for meta-analysis. In the second stage, a generalized least square method is applied after merging and organizing the results from cohorts, to obtain the final estimate of the meta-analysis. The association between any single or multiple haplotypes and the trait can be easily tested, based on our framework.

We evaluate the type-I error rate in a variety of scenarios with different between and within cohort variation. All the scenarios have the correct type-I error rate. We also compare the power of our approach with the univariate min P method adjusted for multiple testing, and demonstrate our approach is at least as powerful and much more powerful in certain scenarios.

Our approach can not only serve as a tool for the discovery of novel associated variants and novel associated regions, Unlike the single-variant and gene-based tests implemented in two separate models, we test the single haplotype effect and the overall effect in one model. it also serves as a complementary tool to single-variant and gene-based tests. From the real application to *G6PC2* region based on the exome-chip project, we find all the top haplotypes built from rare variants are consistent with the single-variant association test results. Moreover, the global test of haplotype association effects is slightly more significant than both SKAT and burden test.

Chapter 5

Summary and Future Work

In this dissertation, we investigate three topics to conduct genetic association tests in family samples. In the first topic, we propose a novel approach to test the association between a genetic variant and a multinomial trait in family samples. Examples of multinomial traits include a three-category variable based on T2D status and obesity: diabetic and obese, diabetic and non-obese and non-diabetic. We test the association of the three classes in one model instead of performing two association tests. Moreover, because there is no ordinal trend in the three classes, it is more reasonable to use a multinomial model instead of an ordinal model.

Our approach is efficient in terms of conducting large-scale association studies. We estimate the variance component only once in the first stage using the phenotype and the covariates, and then in the second stage we test the association between the genetic variants and the trait treating the variance component as fixed. Our approach has the correct type-I error rate in the scenarios evaluated and is shown to be more powerful than GLMM in certain scenarios. We apply our newly developed approach to genetic variants on chromosome 16 using FHS SHARe data and three-class obesity status using

FHS phenotype dataset. We not only replicate the association with the obesity gene *FTO*, but also identify other obesity genes and gene associated with other metabolic traits. Although we currently assume a canonical link function for the multinomial model, we will generalize it to any applicable link functions in the future.

In the second topic, we develop a novel approach to test the association between a genetic variant and bivariate phenotypes in family samples. Based on EGEE, we successfully incorporate the correlation parameters into the estimation framework for regression parameters. Our approach is efficient and stable. Unlike GEE, we combine the modeling of the overall variance of the bivariate phenotype with the use of the kinship matrix, so that we can estimate the regression parameters and correlation parameters simultaneously. Currently, we are using the second-order taylor expansion to approximate the variance and any pairwise covariance, which has more precision than the delta method. In the future, we want to explore the use of higher-order taylor expansion and to evaluate the potential enhancement in terms of precision. In the simulation studies, we calculate both Wald and score test statistics for several MAF scenarios ranging from 0.005 to 0.3, and conclude Wald test yields the correct type-I error rate for common variants (MAF$>= 5\%$) while score test yields the correct type-I error rate for low-frequency variants (MAF$< 5\%$). In the data analysis section, we apply our approach to study the association between the genetic variants on chromosome 16 of FHS SHARe data and the bivariate phenotypes BMI and T2D status, because the obesity gene *FTO* is also known to be associated with T2D status [6] [47] [55] [56]. Not surprisingly, 15 variants out of our top 20 variants are on *FTO* gene , a gene known for its strong association with BMI. The other top 5 variants identified are on gene *ADCY9* which is also known to be associated with BMI [40]. We currently assume no dispersion for the binary trait. In the future, we want to consider the possibility of having overdispersion, because it very commonly

occurs in binary data. We also want to develop a model for the bivariate phenotypes when one phenotype is count data, like CD4 counts.

In the third topic, we develop a general meta-analysis approach for haplotype association tests. We put no restrictions on the haplotypes observed from each cohort, so that we can meta-analyze results from any number of cohorts. Our approach consists of two stages. In the first stage, we conduct the haplotype association test at the cohort level, allowing for familial correlation when appropriate. We regress the phenotype on the expected haplotype dosage conditional on the genotypes while adjusting for covariates. In the second stage, based on the estimates of the regression parameters and the covariance matrix returned by each cohort, we implement a weighted least square method to obtain the haplotype effect estimates of the meta-analysis. Our simulation studies show that our approach has the correct type-I error rate in the scenarios evaluated even when the between-cohort variation is large. We apply our approach to a known region in an glycemia-T2D exome-chip project, and the global test of the haplotype effects is even more significant than the corresponding gene-based and single-variant tests. Our current approach assumes that all SNPs are available in all cohorts. In the future, I will continue to study the situation when some SNPs are not available in some cohorts.

Appendix

1. The covariance between the quantitative and the binary traits, for the same subject is written as

$$cov(Y_{ija}, Y_{ijb}) = E[cov(Y_{ija}, Y_{ijb})|b] + cov(E(Y_{ija}|b), E(Y_{ijb}|b)) \tag{5.1}$$

because

$cov(Y_{ija}, Y_{ijb}|b_{01}, b_{02}) = r\sqrt{var(Y_{ija}|b_{01})var(Y_{ijb}|b_{02})} = r\sqrt{\sigma_e^2\mu_{ij2}(1-\mu_{ij2})}$

$\mu_{ij2} = logit^{-1}(X_{ij}^T\beta_2 + b_{0ij2})$

$E[cov(Y_{ija}, Y_{ijb})|b] = E[r\sqrt{\sigma_e^2\mu_{ij2}(1-\mu_{ij2})}] = r\sigma_e E[\sqrt{\frac{e^{X_{ij}^T\beta_2+b_{02ij}}}{(1+e^{X_{ij}^T\beta_2+b_{02ij}})^2}}]$

After applying second-order taylor expansion with respect to b_{02ij}, we have:

$\approx r\sigma_e\left(\frac{e^{X_{ij}^T\beta_2/2}}{1+e^{X_{ij}^T\beta_2}} + \frac{1}{16}\frac{e^{X_{ij}^T\beta_2/2}(1+e^{X_{ij}^T\beta_2})(1-6e^{X_{ij}^T\beta_2}+e^{2X_{ij}^T\beta_2})}{(1+e^{X_{ij}^T\beta_2})^4}\sigma_2^2\right)$

$cov(E(Y_{ija}|b), E(Y_{ijb}|b)) = 0;$

then we have

$$cov(Y_{ija}, Y_{ijb}) = r\sigma_e\left(\frac{e^{X_{ij}^T\beta_2/2}}{1+e^{X_{ij}^T\beta_2}} + \frac{1}{16}\frac{e^{X_{ij}^T\beta_2/2}(1-6e^{X_{ij}^T\beta_2}+e^{2X_{ij}^T\beta_2})}{\left(1+e^{X_{ij}^T\beta_2}\right)^3}\sigma_2^2\right). \tag{5.2}$$

2. The covariance between the quantitative and the binary traits, for two subjects in the

same family ($\forall j' \neq j$)

$$cov(Y_{ija}, Y_{ij'b}) = E[cov(Y_{ija}, Y_{ij'b})|b] + cov(E(Y_{ija}|b), E(Y_{ij'b}|b)) \tag{5.3}$$

because

$cov(Y_{ija}, Y_{ij'b}|b_{01}, b_{02}) = r_{jj'}\sqrt{var(Y_{ija}|b_{01})var(Y_{ij'b}|b_{02})} = r_{jj'}\sqrt{\sigma_e^2\mu_{ij'2}(1-\mu_{ij'2})}$

$$cov(Y_{ija}, Y_{ij'b}) = r_{jj'}\sigma_e\left(\frac{e^{X_{ij'}^T\beta_2/2}}{1+e^{X_{ij'}^T\beta_2}} + \frac{1}{16}\frac{e^{X_{ij'}^T\beta_2/2}(1-6e^{X_{ij'}^T\beta_2}+e^{2X_{ij'}^T\beta_2})}{\left(1+e^{X_{ij'}^T\beta_2}\right)^3}\sigma_2^2\right) \tag{5.4}$$

.

3. variance-covariance of the continuous trait

   For the same subject:

$$var(Y_{ija}) = \sigma_1^2(\Sigma_{kin})_{ij,ij} + \sigma_e^2. \tag{5.5}$$

   For two subjects in the same family ($j \neq j'$):

$$cov(Y_{ija}, Y_{ij'a}) = \sigma_1^2(\Sigma_{kin})_{ij,ij'}. \tag{5.6}$$

4. variance-covariance of the binomial trait

   For the same subject:

$$var(Y_{ijb}) = E[var(Y_{ijb}|b)] + var(E[Y_{ijb}|b]) \tag{5.7}$$

because

$$E[var(Y_{ijb}|b)] = E[\mu_{ijb}(1-\mu_{ijb})] \tag{5.8}$$

$$var(E[Y_{ijb}|b]) = var(\mu_{ijb}) = E[\mu_{ijb}^2] - E^2[\mu_{ijb}] \tag{5.9}$$

thus

$$var(Y_{ijb}) = E[\mu_{ijb}] - E^2[\mu_{ijb}] \tag{5.10}$$

$$var(Y_{ijb}) = \left(\frac{e^{X_{ij}^T\beta_2}}{1+e^{X_{ij}^T\beta_2}} + \frac{e^{X_{ij}^T\beta_2}\left(1-e^{X_{ij}^T\beta_2}\right)}{4\left(1+e^{X_{ij}^T\beta_2}\right)^3}\sigma_2^2\right)\left(\frac{1}{1+e^{X_{ij}^T\beta_2}} - \frac{e^{X_{ij}^T\beta_2}\left(1-e^{X_{ij}^T\beta_2}\right)}{4\left(1+e^{X_{ij}^T\beta_2}\right)^3}\sigma_2^2\right). \tag{5.11}$$

For two subjects in the same family:

$$cov(Y_{ijb}, Y_{ij'b}) = E[Y_{ijb}Y_{ij'b}] - E[Y_{ijb}]E[Y_{ij'b}]. \tag{5.12}$$

Given conditional independence, we have:

$$\begin{aligned} cov(Y_{ijb}, Y_{ij'b}) &= E[E[Y_{ijb}Y_{ij'b}|b]] - E[E[Y_{ijb}|b]]E[E[Y_{ij'b}|b]] \\ &= E[E[Y_{ijb}|b]E[Y_{ij'b}|b]] - E[E[Y_{ijb}|b]]E[E[Y_{ij'b}|b]] \\ &= E[\mu_{ijb}\mu_{ij'b}] - E[\mu_{ijb}]E[\mu_{ij'b}] \end{aligned} \tag{5.13}$$

$$cov(Y_{ijb}, Y_{ij'b}) = \frac{e^{(X_{ij}+X_{ij'})^T\beta_2}}{(1+e^{X_{ij}^T\beta_2})^2(1+e^{X_{ij'}^T\beta_2})^2}\left[\sigma_2^2(\Sigma_{kin})_{ij,ij'} - \frac{(1-e^{X_{ij}^T\beta_2})(1-e^{X_{ij'}^T\beta_2})\sigma_2^4}{16(1+e^{X_{ij}^T\beta_2})(1+e^{X_{ij'}^T\beta_2})}\right]. \tag{5.14}$$

where $\mu_{ijb} = \dfrac{e^{X_{ij}\beta_2 + b_{02ij}}}{1+e^{X_{ij}\beta_2+b_{02ij}}}$ and $\mu_{ij'b} = \dfrac{e^{X_{ij'}\beta_2+b_{02ij'}}}{1+e^{X_{ij'}\beta_2+b_{02ij'}}}$

Bibliography

[1] Mordecai Avriel. *Nonlinear programming: analysis and methods*. Courier Dover Publications, 2003.

[2] Liana K Billings and Jose C Florez. The genetics of type 2 diabetes: what have we learned from gwas? *Annals of the New York Academy of Sciences*, 1212(1):59–77, 2010.

[3] Joseph-Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer, 2006.

[4] Han Chen, James B Meigs, and Josée Dupuis. Sequence kernel association test for quantitative traits in family samples. *Genetic epidemiology*, 37(2):196–204, 2013.

[5] Wei-Min Chen, Ani Manichaikul, and Stephen S Rich. A generalized family-based association test for dichotomous traits. *The American Journal of Human Genetics*, 85(3):364–376, 2009.

[6] Yoon Shin Cho, Chien-Hsiun Chen, Cheng Hu, Jirong Long, Rick Twee Hee Ong, Xueling Sim, Fumihiko Takeuchi, Ying Wu, Min Jin Go, Toshimasa Yamauchi, et al. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east asians. *Nature genetics*, 44(1):67–72, 2012.

[7] Chia-Min Chung, Tsung-Hsien Lin, Jaw-Wen Chen, Hsin-Bang Leu, Hsin-Chou Yang, Hung-Yun Ho, Chih-Tai Ting, Sheng-Hsiung Sheu, Wei-Chuan Tsai, Jyh-Hong Chen, et al. A genome-wide association study reveals a quantitative trait locus of adiponectin on cdh13 that predicts cardiometabolic outcomes. *Diabetes*, 60(9):2417–2423, 2011.

[8] Zari Dastani, Marie-France Hivert, Nicholas Timpson, John RB Perry, Xin Yuan, Robert A Scott, Peter Henneman, Iris M Heid, Jorge R Kizer, Leo-Pekka Lyytikäinen, et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS genetics*, 8(3):e1002607, 2012.

[9] G Diao and DY Lin. Variance-components methods for linkage and association analysis of ordinal traits in general pedigrees. *Genetic epidemiology*, 34(3):232–237, 2010.

[10] Josée Dupuis, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U Jackson, Eleanor Wheeler, Nicole L Glazer, Nabila Bouatia-Naji, Anna L Gloyn, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105–116, 2010.

[11] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[12] Timothy M Frayling, Nicholas J Timpson, Michael N Weedon, Eleftheria Zeggini, Rachel M Freathy, Cecilia M Lindgren, John RB Perry, Katherine S Elliott, Hana Lango, Nigel W Rayner, et al. A common variant in the fto gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889–894, 2007.

[13] Xiaoyi Gao, Joshua Starmer, and Eden R Martin. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic epidemiology*, 32(4):361–369, 2008.

[14] Jelle Goeman, Jan Oosting, Maintainer Jelle Goeman, and Imports Biobase. Package 'globaltest'. 2012.

[15] Daniel B Hall. On the application of extended quasi-likelihood to the clustered data case. *Canadian Journal of Statistics*, 29(1):77–97, 2001.

[16] Daniel B Hall and Thomas A Severini. Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association*, 93(444):1365–1375, 1998.

[17] Audrey E Hendricks, Josée Dupuis, Mark W Logue, Richard H Myers, and Kathryn L Lunetta. Correction for multiple testing in a gene region. *European Journal of Human Genetics*, 22(3):414–418, 2014.

[18] Kikuko Hotta, Takuya Kitamoto, Aya Kitamoto, Seiho Mizusawa, Tomoaki Matsuo, Yoshio Nakata, Seika Kamohara, Nobuyuki Miyatake, Kazuaki Kotani, Ryoya Komatsu, et al. Association of variations in the fto, scg3 and mtmr9 genes with metabolic syndrome in a japanese population. *Journal of human genetics*, 56(9):647–651, 2011.

[19] Kikuko Hotta, Yoshio Nakata, Tomoaki Matsuo, Seika Kamohara, Kazuaki Kotani, Ryoya Komatsu, Naoto Itoh, Ikuo Mineo, Jun Wada, Hiroaki Masuzaki, et al. Variations in the fto gene are associated with severe obesity in the japanese. *Journal of human genetics*, 53(6):546–553, 2008.

[20] Yi-Juan Hu, Sonja I Berndt, Stefan Gustafsson, Andrea Ganna, Joel Hirschhorn, Kari E North, Erik Ingelsson, and Dan-Yu Lin. Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *The American Journal of Human Genetics*, 93(2):236–248, 2013.

[21] Sun Ha Jee, Jae Woong Sull, Jong-Eun Lee, Chol Shin, Jongkeun Park, Heejin Kimm, Eun-Young Cho, Eun-Soon Shin, Ji Eun Yun, Ji Wan Park, et al. Adiponectin concentrations: a genome-wide association study. *The American Journal of Human Genetics*, 87(4):545–552, 2010.

[22] J Li and L Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227, 2005.

[23] Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

[24] Dajiang J Liu, Gina M Peloso, Xiaowei Zhan, Oddgeir L Holmen, Matthew Zawistowski, Shuang Feng, Majid Nikpay, Paul L Auer, Anuj Goel, He Zhang, et al. Meta-analysis of gene-level tests for rare variant association. *Nature genetics*, 46(2):200–204, 2014.

[25] Jianfeng Liu, Yufang Pei, Chris J Papasian, and Hong-Wen Deng. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genetic epidemiology*, 33(3):217–227, 2009.

[26] Ani Manichaikul, Wei-Min Chen, Kayleen Williams, Quenna Wong, Michèle M Sale, James S Pankow, Michael Y Tsai, Jerome I Rotter, Stephen S Rich, and

Josyf C Mychaleckyj. Analysis of family-and population-based samples in cohort genome-wide association studies. *Human genetics*, 131(2):275–287, 2012.

[27] Alisa K Manning, Michael LaValley, Ching-Ti Liu, Kenneth Rice, Ping An, Yongmei Liu, Iva Miljkovic, Laura Rasmussen-Torvik, Tamara B Harris, Michael A Province, et al. Meta-analysis of gene-environment interaction: joint estimation of snp and snp× environment regression coefficients. *Genetic epidemiology*, 35(1):11–18, 2011.

[28] M Marra, F Pasanisi, L Scalfi, P Colicchio, M Chelucci, and F Contaldo. The prediction of basal metabolic rate in young adult, severely obese patients using single-frequency bioimpedance analysis. *Acta diabetologica*, 40(1):s139–s141, 2003.

[29] Peter McCullagh and John A Nelder. Generalized linear models. 1989.

[30] Hiroko Morisaki, Itaru Yamanaka, Naoharu Iwai, Yoshihiro Miyamoto, Yoshihiro Kokubo, Tomonori Okamura, Akira Okayama, and Takayuki Morisaki. Cdh13 gene coding t-cadherin influences variations in plasma adiponectin levels in the japanese population. *Human mutation*, 33(2):402–410, 2012.

[31] Dale R Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765–769, 2004.

[32] Peter C O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*, pages 1079–1087, 1984.

[33] Yukiyoshi Okauchi, Ken Kishida, Tohru Funahashi, Midori Noguchi, Tomoko Ogawa, Miwa Ryo, Kohei Okita, Hiromi Iwahashi, Akihisa Imagawa, Tadashi Nakamura, et al. Changes in serum adiponectin concentrations correlate with changes

in bmi, waist circumference, and estimated visceral fat area in middle-aged general population. *Diabetes Care*, 32(10):e122–e122, 2009.

[34] Paul F O'Reilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and Lachlan JM Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PLoS One*, 7(5):e34861, 2012.

[35] Karim Oualkacha, Zari Dastani, Rui Li, Pablo E Cingolani, Timothy D Spector, Christopher J Hammond, J Brent Richards, Antonio Ciampi, and Celia MT Greenwood. Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic epidemiology*, 37(4):366–376, 2013.

[36] Angelo Pietrobelli, Robert C Lee, Esmeralda Capristo, Richard J Deckelbaum, and Steven B Heymsfield. An independent, inverse association of high-density-lipoprotein-cholesterol concentration with nonadipose body mass. *The American journal of clinical nutrition*, 69(4):614–620, 1999.

[37] José C Pinheiro and Douglas M Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, 4(1):12–35, 1995.

[38] María E Sáez, Antonio González-Pérez, María T Martínez-Larrad, Javier Gayán, Luis M Real, Manuel Serrano-Ríos, and Agustín Ruiz. Wwox gene is associated with hdl cholesterol and triglyceride levels. *BMC medical genetics*, 11(1):148, 2010.

[39] Robert A Scott, Vasiliki Lagou, Ryan P Welch, Eleanor Wheeler, May E Montasser, Jian'an Luan, Reedik Mägi, Rona J Strawbridge, Emil Rehnberg, Stefan Gustafsson, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics*, 44(9):991–1005, 2012.

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

[41] Richard S Spielman and Warren J Ewens. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *The American Journal of Human Genetics*, 62(2):450–458, 1998.

[42] Richard S Spielman, Ralph E McGinnis, and Warren J Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American journal of human genetics*, 52(3):506, 1993.

[43] Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PloS one*, 8(7):e65245, 2013.

[44] Terry Therneau. *coxme: Mixed Effects Cox Models.*, 2012. R package version 2.2-3.

[45] Sophie van der Sluis, Danielle Posthuma, and Conor V Dolan. Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS genetics*, 9(1):e1003235, 2013.

[46] DJ Venzon and SH Moolgavkar. A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, pages 87–94, 1988.

[47] Benjamin F Voight, Laura J Scott, Valgerdur Steinthorsdottir, Andrew P Morris, Christian Dina, Ryan P Welch, Eleftheria Zeggini, Cornelia Huth, Yurii S Aulchenko, Gudmar Thorleifsson, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics*, 42(7):579–589, 2010.

[48] Xueqin Wang, Yuanqing Ye, and Heping Zhang. Family-based association tests for ordinal traits adjusting for covariates. *Genetic epidemiology*, 30(8):728–736, 2006.

[49] Jennifer Wessel, Audrey Y Chu, Sara M Willems, Shuai Wang, Hanieh Yaghootkar, Jennifer A Brody, Marco Dauriz, Marie-France Hivert, Sridharan Raghavan, Leonard Lipovich, et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nature communications*, 6, 2015.

[50] Ying Wu, Yun Li, Ethan M Lange, Damien C Croteau-Chonka, Christopher W Kuzawa, Thomas W McDade, Li Qin, Ghenadie Curocichin, Judith B Borja, Leslie A Lange, et al. Genome-wide association study for adiponectin levels in filipino women identifies cdh13 and a novel uncommon haplotype at kng1–adipoq. *Human molecular genetics*, page ddq423, 2010.

[51] Hsin-Chou Yang, Yu-Jen Liang, Jaw-Wen Chen, Kuang-Mao Chiang, Chia-Min Chung, Hung-Yun Ho, Chih-Tai Ting, Tsung-Hsien Lin, Sheng-Hsiung Sheu, Wei-Chuan Tsai, et al. Identification of igf1, slc4a4, wwox, and sfmbt1 as hypertension susceptibility genes in han chinese with a genome-wide gene-based association study. *PloS one*, 7(3):e32907, 2012.

[52] Qiong Yang, Hongsheng Wu, Chao-Yu Guo, and Caroline S Fox. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic epidemiology*, 34(5):444–454, 2010.

[53] Dmitri V Zaykin, Peter H Westfall, S Stanley Young, Maha A Karnoub, Michael J Wagner, and Margaret G Ehm. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human heredity*, 53(2):79–91, 2002.

[54] Scott L Zeger, Kung-Yee Liang, and Paul S Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.

[55] Eleftheria Zeggini, Laura J Scott, Richa Saxena, Benjamin F Voight, Jonathan L Marchini, Tianle Hu, Paul IW de Bakker, Gonçalo R Abecasis, Peter Almgren, Gitte Andersen, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40(5):638–645, 2008.

[56] Eleftheria Zeggini, Michael N Weedon, Cecilia M Lindgren, Timothy M Frayling, Katherine S Elliott, Hana Lango, Nicholas J Timpson, John RB Perry, Nigel W Rayner, Rachel M Freathy, et al. Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341, 2007.

Shuai Wang

801 Massachusetts Avenue, 3rd Floor

Boston, MA 02118

tutuwang@bu.edu

Education

Ph.D. in Biostatistics, ***Boston University*** *May 2015*

M.S. in Statistics, ***University of Virginia***, VA *May 2011*

B.S. in Mathematics and Applied Mathematics, ***Fudan University***, China *Sep 2009*

Programming Skills

Proficient in: R, SAS9.2, Python, Perl, OpenBUGS, SQL, Shell Scripting, MATLAB, FORTRAN, C++

Software: STATA, MACH, Merlin, PLINK/SEQ, RAREMETAL

Operating System: UNIX Computing Cluster, Windows

Research Interests

Statistical Genetics, Correlated Data Analysis, Large-Scale Computing (Big Data), Data Mining

Industry Experience

Summer Intern, ***AstraZeneca***, Waltham, MA *May 2014-Aug 2014*

Summer Statistics Intern, ***Liberty Mutual***, Boston, MA *May 2012-Aug 2012*

Statistical Genetics Intern, ***Biogen Idec***, Cambridge, MA *Sep 2011-May 2012*

Research Experience

- Research Assistant, ***Framingham Heart Study***(FHS), BUSPH, Boston, MA

  Aug 2012- Present

  Leading analyst in the glycemic-T2D working group

- Research Assistant, ***Center of Excellence in Sickle Cell Disease***, BUMC, Boston

  Aug 2012- Aug 2013

  Worked on the whole-genome sequencing project of saudi-arabian sickle cell diseased patients

- Research Assistant, ***Center for Economics and Public Policy***, UVa

  Feb 2010- May 2011

  Maintained the database of the Center in MySQL and provided statistical consulting to various economic and public policy projects

- Statistical Consulting, Supervised by Prof. Keenan, Department of Statistics, UVa

  Spring 2010

  Worked on T2D dose response model

Teaching

Department of Biostatistics, ***BUSPH***, Boston, MA *2012-Present*

- Instructor, BS723 Introduction to Statistical Computing *Spring 2015*
- TA, BS855 Bayesian Modeling for Biomedical Research & Public Health *Fall 2014*
- TA, BS720 Introduction to R *Spring 2014*
- TA, BS858 Statistical Genetics I *Fall 2013*
- TA, BS723 Introduction to Statistical Computing (SAS) *Summer 2012*

Publications

1. Wang S, Fisher V, Chen Y, Dupuis J. Comparison of multi-SNV association tests in a meta-analysis of GAW19 family and unrelated data. *Accepted* BMC Proceedings 2015

2. Sebastiani P, Farrell J, Alsultan A, Wang S, Edward H, Shappell H, Bae H, et al. BCL11A Enhancer Haplotypes and Fetal Hemoglobin in Sickle Cell Anemia. *Accepted*. Blood Cells, Molecules and Disease

3. Wang S, Hu F, Dupuis J. Genetics of Type 2 Diabetes and Related Traits. Springer. Book Chapter. *To Appear*

4. Wessel J*, Chu A*, Willems S*, Wang S*, et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nature Communications. 2015 Jan; 6:5897. DOI: 10.1038/ncomms6897. *: Equally Contributed

5. Wang S, Gao W, Ngwa J, Allard C, Liu CT, Cupples LA. Comparing baseline and longitudinal measures in association studies. BMC Proceedings 2014, 8(Suppl 1):S84 doi:10.1186/1753-6561-8-S1-S84

6. Cornes B, Brody J, NIKPOOR N, Morrison A, Wang S, et al. Association of Levels of Fasting Glucose and Insulin with Rare Variants at the Chromosome 11p11.2 MADD Locus: the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Targeted Sequencing Study. Circ Cardiovasc Genet. 2014 Jun;7(3):374-82. doi: 10.1161/CIRCGENETICS.113.000169

Presentations

1. CHARGE Investigator Meeting. Nov 11-13, 2014. Washington, DC. Presentation Title: Gene-based Tests with Defined Functional Categories Revealed Novel Findings

2. Genetic Analysis Workshop 19 (GAW19). Aug 24-27, 2014. Vienna, Austria

3. International Genetic Epidemiology Society (IGES). Aug 28-30, 2014. Vienna, Austria, *Williams Award Finalist* (2 pre-doctoral students selected from over 150 abstracts). Presentation Title: Meta-analysis approach for haplotype association tests: a general framework for family and unrelated samples

4. Joint Statistical Meeting (JSM). Aug 2-7, 2014. Boston, MA. Presentation Title: A General Meta-Analysis Approach for Haplotype Association Results in Family and Unrelated Samples

5. 2014 Winter CHARGE Investigator Meeting. Jan 22-24, 2014. Redondo Beach, CA. Poster Title: Haplotype association Analysis of the G6PC2 Region with Fasting Glucose level

Academic Awards

NIH Travel Award, November 11-13 CHARGE meeting in Washington, D.C. *Sep 2014*

GAW19 Travel Award, August 24-27, Genetic Analysis Workshop in Vienna, Austria *Jun 2014*

NIH Travel Award, January 22-24 CHARGE Meeting in Redondo Beach, CA *Nov 2013*

Academic Award, Department of Statistics, University of Virginia *May 2011*

Outstanding Graduates, Fudan University, Shanghai, China *Jun 2009*

A-level Graduation Thesis, Fudan University, Shanghai, China *Jun 2009*

Third-class Scholarship, Fudan University, Shanghai, China *2006-2008*

Credentials

SAS Certified Advanced Programmer for SAS9.2, 'Probability'(10) of SOA

Membership

American Statistical Association (ASA) *2012-Present*