2021

Exploiting family history in genetic analysis of rare variants

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

EXPLOITING FAMILY HISTORY IN GENETIC ANALYSIS OF RARE VARIANTS

by

YANBING WANG

B.S., Kent State University, 2015
M.S., Brown University, 2017

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2021

Approved by

First Reader ____

Josée Dupuis, Ph.D.
Professor and Chair of Biostatistics
Boston University, School of Public Health

Second Reader ____

Gina Marie Peloso, Ph.D.
Associate Professor of Biostatistics
Boston University, School of Public Health

Third Reader ____

Anita L. DeStefano, Ph.D.
Professor of Biostatistics
Boston University, School of Public Health

Associate Professor of Neurology
Boston University, School of Medicine

Fourth Reader ____

Chunyu Liu Ph.D.
Associate Professor of Biostatistics
Boston University, School of Public Health

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my thesis advisor Dr. Josée Dupuis, who continuously guided and supported me throughout my PhD journey, and who shown me what a good research scientist and positive person should be. Thanks for journeying with me and sharing the excitement of discovery along the way. I owe a debt of gratitude for her time, ethnicism, patience, and knowledge. She has taught and inspired me more than I could ever give the credit for her here. This work could not have been possible without her.

I would like to thank Dr. Anita DeStefano for giving me the opportunity working in the Alzheimer's Disease Sequencing Project (ADSP) and bringing me to the field of statistical genetics. Her training and supervision in aspects of doing good research are greatly appreciated. Without her support, I could not have better balanced my school and family. I would like to thank Dr. Gina Peloso and Dr. Han Chen for always providing insightful and helpful comments on my work. My thanks also go to Dr. Ching-Ti Liu and Dr. Chunyu Liu for their advice and constructive criticism in my dissertation.

I am also grateful to those with whom I have had the pleasure to work in the ADSP group: Dr. Honghuang Lin, Dr. Chloe Sarnowski, Dr. Nancy Heard-Costa, and Dr. Achilleas Pitsillides, for their support in my research. My appreciation also extends to my friends at Boston University who made my PhD study more enjoyable.

Lastly, I would especially thank with love to my family. My parents and my sister, whose unconditional love, encouragement, and support are always with me in whatever I pursue. My husband, who has never stopped encouraging me towards success and has been

very supportive of me during this process. Thank you my daughter, you made me stronger, fulfilled, and better than I could have ever imagined.

EXPLOITING FAMILY HISTORY IN

GENETIC ANALYSIS OF RARE VARIANTS

YANBING WANG

Boston University Graduate School of Arts and Sciences, 2021

Major Professor: Josée Dupuis, Professor and Chair of Biostatistics, Boston University, School of Public Health

ABSTRACT

Genetic association analyses have successfully identified thousands of genetic variants contributing to complex disease susceptibility. However, these discoveries do not explain the full heritability of many diseases, due to the limited statistical power to detect loci with small effects, especially in regions with rare variants. The development of new and powerful methods is necessary to fully characterize the underlying genetic basis of complex diseases. Family history (FH) contains information on the disease status of un-genotyped relatives, which is related to the genotypes of probands at disease loci. Exploiting available FH in relatives could potentially enhance the ability to identify associations by increasing sample size. Many studies have very low power for genetic research in late-onset diseases because younger participants do not contribute a sufficient number of cases and older patients are more likely deceased without genotypes. Genetic association studies relying on cases and controls need to progress by incorporating additional information from FH to expand genetic research.

This dissertation overcomes these challenges and opens up a new paradigm in genetic research. The first chapter summarizes relevant methods used in this dissertation. In the second chapter, we develop novel methods to exploit the availability of FH in

aggregation unit-based test, which have greater power than other existing methods that do not incorporate FH, while maintaining a correct type I error. In the third chapter, we develop methods to exploit FH while adjusting for relatedness using the generalized linear mixed effect models. Such adjustment allows the methods to have well-controlled type I error and maintain the highest sample size because there is no need to restrict the analysis to an unrelated subset in family studies. We demonstrate the flexibility and validity of the methods to incorporate FH from various relatives. The methods presented in the fourth chapter overcome the issue of inflated type I error caused by extremely unbalanced case-control ratio. We propose robust versions of the methods developed in the second and third chapters, which can provide more accurate results for unbalanced study designs. Availability of these novel methods will facilitate the identification of rare variants associated with complex traits.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACAT	Aggregated Cauchy Association Test
ACAT-V	Aggregated Cauchy Association Test – Variant-set Test
AD	Alzheimer's Disease
BMI	Body Mass Index
BOLT-LMM	Bayesian Mixed-Model Association Method
CGF	Cumulant Generating Function
cumMAC	Cumulative Minor Allele Count
ER	Efficient Resampling
famBT	Family-based Burden Test
famFHAT	Family-based Family History Aggregation unit-based Test
famFHAT-O	Family-based Family History Aggregation unit-base Test – Optimal Test
famFHAT-Burden	Family-based Family History Aggregation unit-based Test - Burden Test
famSKAT	Family-based Sequence Kernel Association Test
famSKAT-O	Family-based Sequence Kernel Association Test – Optimal Test
FE	Functionally Equivalent
FH	Family History
FHAT	Family History Aggregation unit-based Test

FHAT-O	Family History Aggregation unit-based Test – Optimal Test
FHAT-Burden	Burden Family History Aggregation unit-based Test – Burden Test
FHS	Framingham Heart Study
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Effect Model
GMMAT	Generalized linear Mixed Model Association Tests
GWAS	Genome-Wide Association Studies
GWAX	Genome-Wide Association by Proxy
LD	Linkage Disequilibrium
LT-FH	Liability Threshold model of case-control status and Family History
MAC	Minor Allele Count
MAF	Minor Allele Frequency
MiST	Mixed effects Score Test
MONSTER	Minimum P-value Optimized Nuisance parameter Score Test Extended to Relatives
PC	Principal Component
QC	Quality Control
robust-famFHAT	Robust version of Family-based Family History Aggregation unit-based Test

robust-famFHAT-O	Robust version of Family-based Family History Aggregation unit-based Test – Optimal Test
robust-famSKAT	Robust version of Family-based Sequence Kernel Association Test
robust-famSKAT-O	Robust version of Family-based Sequence Kernel Association Test – Optimal Test
robust-FHAT	Robust version of Family History Aggregation unit-based Test
robust-FHAT-O	Robust version of Family History Aggregation unit-based Test – Optimal Test
robust-SKAT	Robust version of Sequence Kernel Association Test
robust-SKAT-O	Robust version of Sequence Kernel Association Test – Optimal Test
SKAT	Sequence Kernel Association Test
SKAT-LTFH	Sequence Kernel Association Test using Liability Threshold model of case-control status and Family History phenotypes
SKAT-O	Sequence Kernel Association Test – Optimal Test
SKATO-LTFH	Sequence Kernel Association Test using Liability Threshold model of case-control status and Family History phenotypes – Optimal Test
SMMAT	Variant Set Mixed Model Association Tests

SMMAT-E	Variant Set Mixed Model Association Tests – Combination of Burden Test and SKAT
SPA	Saddle Point Approximation
T2D	Type 2 Diabetes
VEP	Ensemble Variant Effect Predictor

Chapter 1 Introduction

1.1 Background

Genetic association analysis is often used to evaluate the association between genetic variants and a particular disease in individuals who have the disease (cases) and individuals who don't have the disease (controls), to identify disease-related loci. In the presence of association, disease-susceptible genetic variants will be observed more often in cases. Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex diseases at a genome-wide significance level ($P < 5x10^{-8}$). Most of the variants identified by GWAS are common variants with minor allele frequency (MAF) ≥ 1%, and most of these variants display modest effect sizes and can only explain a small portion of the total heritability of complex diseases. [16] The investigation of rare variants (MAF < 1%) that contribute to the unexplained heritability may lead to the identification of novel genes related to complex traits. However, the standard association tests evaluating each variant individually are grossly underpowered for rare variants. [38, 36, 32] In the last decade, many methods to jointly analyze variants within an aggregation unit (i.e., gene) have been proposed, and these methods have demonstrated improved power to detect rare variant associations. [38, 59, 48, 49, 42, 37, 40] However, power of these methods to identify susceptibility regions for complex diseases is often limited by an insufficient number of cases in unascertained cohorts.

Family history (FH) information provides an overview of phenotypes within families. Such information typically includes phenotypes of un-genotyped parents and may likely capture more distant relatives of probands. For heritable diseases, the relatives' phenotypes will be related to the probands' genotypes at disease loci based on the Mendelian laws of transmission, adding valuable information about the probands' health and risk of diseases. [53, 55, 56] In genetic association studies, the FH for various diseases is often collected in large population cohorts. [17, 14] While collecting cases is expensive, incorporating the FH of the disease into the standard case-control genetic association analysis is a cost-effective way to potentially increase statistical power to detect associations. [19, 39, 43]

For late-onset diseases such as Alzheimer's disease (AD), patients may be deceased with unavailable genotype data, limiting the use of study designs such as trio family-based studies for genetic research. Moreover, the prevalence of some diseases in younger cohorts is very low, and the standard statistical association tests with a small number of cases are not powerful to identify genetic regions associated to a trait of interest. In contrast, the incorporation of the FH of a disease can largely increase the effective sample size in cohorts with limited cases, and shift the current paradigm of genetic research. Therefore, genetic association studies using only cases and controls need to progress by incorporating available FH information to enhance genetic research.

There are a few reported methods exploiting FH in genetic association analysis to improve statistical power to detect disease loci. [19, 39, 20, 26] Nevertheless, these methods suffer from low power to detect rare variant associations. Because rare variants play an important role in understanding unexplained heritability and discovering genes causally related to complex traits, numerous variant-set methods to jointly analyze rare variants have been proposed to improve power to detect rare variant associations. [38, 59, 48, 49, 42, 37, 40] To our knowledge, none of the variant-set methods can directly incorporate FH information.

To understand the genetic etiology of complex diseases, and overcome gaps in statistical methods incorporating FH for rare variant association analysis, we develop powerful aggregation unit-based methods to exploit FH in genetic association analysis. Methods with good statistical power can uncover novel susceptibility genes that are potential targets for disease therapeutics and prevention.

1.2 Leveraging Family History in Single Variant Analysis

Results from a literature survey indicate that FH is usually ignored or not appropriately incorporated in the standard analysis. Ghosh et al. developed a method that enables the incorporation of FH into standard case-control association studies, and the method shows an improved power to detect disease loci in genetic analyses. [19]

More formally, let G^P, Y^P, and X^P denote the disease status, genotype, and covariates of proband, respectively, and let Y^R denote the disease status of a single relative of the proband. In their approach, they assume the disease status of a single ungenotyped parent is available per proband, and the total association analysis is divided into two independent analyses; one is the standard analysis between probands' disease status with their genotypes using the model $logit\left(P(Y^P = 1|G^P)\right) = X^P\alpha_P + G^P\,\beta_P$, where α_P is a coefficient vector for covariate effects, β_P is a coefficient vector for the observed genotypes. The other analysis evaluates the association between parents' disease status and probands' genotypes, conditioning on the proband disease status using the following model: $logit\left(P(Y^R = 1|G^P, X^R)\right) = X^R\alpha_R + G^P\,\beta_R + Y^P\,\lambda_R$. where α_R is a coefficient vector for covariate effects for relatives, β_R is a coefficient vector for the observed genotypes, and λ_R is the coefficient vector for the probands' disease status. The genetic effect estimates ($\hat{\beta}_P$ and $\hat{\beta}_R$) from the two association analyses are asymptotically independent, and the ratio of their expected values is equal to twice the kinship coefficient [55]. Based on these facts, Ghosh et al. suggested to combine the two test statistics with the appropriate weights into a single test that asymptotically follows an standard normal distribution under the null hypothesis.

1.3 Aggregation unit-based Methods

While most of the variants identified by GWAS are common variants with minor allele frequency (MAF) ≥ 1%, they have relatively small effect sizes and can only explain a small portion of the total heritability of complex diseases. [16] The discovery of rare variants (MAF < 1%) that contribute to the unexplained heritability could lead to the identification of novel susceptibility genes. However, the standard association tests evaluating each variant individually are grossly underpowered to detect rare variant associations. [18,33,32] In the last decade, the aggregation unit-based methods to assess the joint effects of multiple variants in a region have been proposed to improved power to detect rare variant associations. [59, 48, 38, 49, 42, 37, 40]

Many aggregation unit-based methods are available for rare variant analysis such as Burden tests [38,42], sequence kernel association test (SKAT) [59], SKAT-O [37], and aggregated Cauchy association test (ACAT) [40]. Burden tests and SKAT are widely used methods because of the computational efficiency and flexibility to incorporate covariates for both quantitative and binary traits. Burden tests are most powerful when all genetic effects are in the same directions and of equal sizes, but suffer from substantial power loss when causal variants have different directions of effects. SKAT is more powerful than Burden tests when variants have different directions of effects on the phenotype. Another key feature for SKAT is that it only requires fitting the null model with covariates once. SKAT-O is an omnibus test combining the features of SKAT and Burden tests, which power is robust when modelling causal variant with both different and same directions of effects. ACAT has shown to outperform SKAT and Burden tests when only a small proportion of variants are associated with the phenotype in a region.

The power of tests depends on the genetic architecture, thus, the choice of weights will affect power. The variants are weighted using user-defined weights in these methods. For example, one can use Wu's weights [59] to up-weight the effects of very rare variants that are believed to have larger effects. The variants can also be weighted based on annotations. [30] Results from variant-set based methods can be meta-analyzed by combining the score statistics from multiple studies. [36] Meta-analysis has shown to be as powerful as the joint analysis by pooling all individuals in the analysis under certain conditions.

1.4 Generalized Linear Mixed Effect Models

The standard genetic association analysis is designed to investigate the variants-disease associations in unrelated samples, where the generalized linear models (GLMs) are widely used. Without accounting for relatedness among samples, the methods would yield inflated type I error in family studies. To deal with this issue, one attractive approach is to use the generalized mixed effect model (GLMM) including random effects through a kinship matrix. With no need to restrict the analysis to unrelated set, this approach yields the higher power than restricting analysis to an unrelated subset as it maximizes the usage of data.

The GLMM can be defined as $g\big(E(Y\,|G\,,X\,,\delta\,)\big) = X\alpha\, + G\beta + \delta$, where $g(\cdot)$ is the link function, Y is the $n \times 1$ phenotype vector, X is the $n \times q$ covariate matrix, G is the $n \times m$ genotype matrix, α is the $q \times 1$ coefficient vector for covariates, and β is the $m \times 1$ coefficient vector for genetic variants. In the model, δ is a random effect that follows the distribution of $N\big(0, \sigma_G^2 \Phi\,\big)$ where Φ_P is twice the kinship matrix calculated using known family information or genetic relationship matrix estimated from genotypes, and σ_G^2 is a variance component parameter. Under the null hypothesis of no genetic effects, for a continuous phenotype with an identity link function, i.e., $Y = X\alpha\, + G\beta + \varepsilon$ with $\varepsilon \sim N(0, \delta_e^2 I)$, the variance-covariance matrix is $\Sigma = \sigma_{G_P}^2 \Phi\, + \delta_e^2 I$.For a binary phenotype with a logit link function, i.e., $logit\,\big(P(Y\, = 1|X,\, G, \delta\,)\big) = X\alpha\, + G\beta + \delta$, $\Sigma = \sigma_{G_P}^2 \Phi\, +$

$diag\left\{\frac{1}{\mu(1-\mu)}\right\}$, where $\mu = E((Y\ = 1|X,\ \delta))$. The covariance that accounts for relatedness among samples can be incorporated in the score test statistics to account for familial correlation in family studies.

1.5 Saddle Point Approximation and Efficient Resampling

Case-control ratios are often extremely unbalanced in large-scale cohorts such as biobanks, due to low disease prevalence. Although normal approximation performs well for the score statistic in balanced study design, the asymptotic assumption of normal distribution is not valid for a binary outcome in the presence of unbalanced case-control ratios. Saddle point approximation (SPA) and efficient resampling (ER) are used to calibrate the variance of the single-variant score statistics, and have been successfully used to address this problem. [60,62]

SPA method approximates the distribution using cumulant generating function (CGF). Let $K_j(T)$ denote the CGF of a single score statistic for variant j. The approximated CGF based on the GLMM can be written as

$$\widehat{K_J}(t;\hat{u},c) = \sum_{i=1}^{N} \log\left(1 - \hat{\mu} + \hat{\mu}e^{ct\tilde{G}}\right) - ct\sum_{i=1}^{N} \tilde{G}\hat{\mu},$$

where $\hat{\mu}$ is phenotype mean estimated from the null GLMM without genetic effects, c is a constant that equals to $Var^{*}(T)^{-1/2}$ with $Var^{*}(T)\ = \tilde{G}W\tilde{G}$ where W is the prespecified weight matrix for the genotypes. The distribution of the score statistic S_j for variant j can be estimated based on the CGF. [62]

1.6 Dissertation Outline

In this dissertation, we develop methods to incorporate family history into analyses of rare genetic variants. In each of the projects, we present the methodological developments, intensive simulation analysis results, and applications to large-scale cohorts including the UK Biobank and Framingham Heart Study (FHS).

In Chapter 2, we develop the family history aggregation unit-based test (FHAT) and optimal FHAT (FHAT-O) that exploit FH information using the GLM. We also demonstrate a novel way to utilize the liability threshold model of case-control status and FH (LT-FH) [26] for rare variant analysis, and compare the performance of our methods to LT-FH. Given that some studies with insufficient cases have very low power for genetic research, FH can be incorporated to greatly increase samples size and enhance the ability to identify genetic associations.

In Chapter 3, we propose the family-based FHAT (famFHAT) and optimal famFHAT (famFHAT-O) to account for family correlation in either probands with available phenotypes and genotypes and relatives with only available phenotypes (i.e., FH) through the GLMM. For the scenarios when family structure is complex, we show that our methods have the flexibility to incorporate FH from multiple relatives with different degrees of relationship with the probands. There is no need to restrict analyses to an unrelated subset thus maintaining a large sample size in most studies.

In Chapter 4, we present the robust versions of the methods developed in Chapters 2 and 3 to allow the analysis of rare variant for studies with extremely unbalanced case-control ratios, where the approaches of SPA and ER are employed to calibrate the variance of score statistics. While offering improved power with the incorporation of FH information, more accurate results under case-control imbalance will be achieved using these robust methods.

In Chapter 5, we summarize our findings and discuss our future work.

Chapter 2 Exploiting Family History in Aggregation Unit-based Genetic Association Tests

2.1 Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex diseases at the genome-wide significance level ($P < 5x10^{-8}$). Most of the variants identified by GWAS are common variants with minor allele frequency (MAF) ≥ 1%, and most of these variants display modest effect sizes and can only explain a small portion of the total heritability of complex diseases. Yet, rare variants (MAF < 1%) are of vital importance to uncovering unexplained heritability and discovering novel genes contributing to complex diseases. [4,16,50] Because standard association approaches testing each variant individually are grossly underpowered for rare variants, methods that jointly analyze variants within aggregation units (i.e., genes) have been proposed to improve power to detect rare variant associations. Aggregation unit-based approaches include, among others, the sequence kernel association test (SKAT) [59], Burden tests [47,38,42], SKAT-O [37], and aggregated Cauchy association test (ACAT) [40]. However, power of these methods to identify disease regions can be limited by insufficient number of cases in unascertained cohorts.

In genetic association studies, family history (FH) of disease in relatives is often collected in large population cohorts. FH provides an overview of phenotypes within families. Such information typically includes phenotypes of un-genotyped parents or more distant relatives of probands. FH is related to the genotypes of probands at disease loci based on the Mendelian laws of transmission, and is important in assessing health problems and risk of diseases. [53,26,56] Many study designs have limitations for genetic research of late-onset diseases such as Alzheimer's disease (AD), because disease cases may be deceased with unavailable genotype data. The standard statistical association tests in younger cohorts with low prevalence of some late-onset diseases are not powerful to identify genetic regions associated to a trait of interest. In contrast, the incorporation of available information of disease status in the form of FH may increase the sample size in cohorts with limited cases or individuals with unavailable genotypes. Genetic association studies using only cases and controls will greatly benefit by incorporating available FH information to detect associations.

FH cannot be directly incorporated in standard genetic association methods, limiting its use in genetic association testing. FH has been included as a covariate to improve disease prediction, [21] or used to infer mode of inheritance to construct statistical tests. [51] However, there are a few reported methods that allow FH to be exploited in genetic association analysis to improve statistical power to detect disease loci. The method developed by Ghosh *et al*. [19] enables the incorporation of FH as a phenotype into the standard single variant analysis, and the results confirmed that exploiting the information contained in FH substantially boosts power to detect the individual variant at disease loci. Nevertheless, these single variant tests suffer from loss of power to detect rare variant associations. While numerous aggregation unit-based methods to jointly analyze rare variants have been proposed to improve power to detect rare variant associations, aggregation unit-based methods that can directly incorporate FH information are needed. In this chapter, we develop a new and powerful method of family history aggregation unit-based test (FHAT) that enables the incorporation of FH to enhance the statistical power for rare variant associations. We also develop an optimal unified test FHAT-O to maintain robust power in complex scenarios regardless of directions of genetic effects or proportions of causal variants. To make the comparison with the recent developed method, liability threshold model of case-control status and FH (LT-FH), [26] we propose a novel way to utilize LT-FH into aggregation unit-based method for rare variant analysis. We perform an extensive simulation study to evaluate the type I error and power of FHAT and FHAT-O under various scenarios. We demonstrate that our methods and the LT-FH method control type I error in a reasonable range of significance levels and relatively better than SKAT, SKAT-O when the disease prevalence is low in probands. With greatly reduced computational cost, FHAT and FHAT-O are more powerful than SKAT-LTFH and SKATO-LTFH when the effect of the variant increases with age. FHAT has greater power than SKAT and ACAT-V in most cases when exploiting additional FH information in relatives, and FHAT-O maintains robust power in various complex scenarios. We conduct the rare variant aggregation unit-based tests using unrelated white participants from the UK Biobank tranche of 200,000 individuals with whole-exome sequencing data for all cause dementia (including AD) and hypertension. With enhanced ability to detect disease susceptibility genetic associations, the new findings enabled by these novel methods will contribute to the understanding of the genetic etiology of complex diseases.

2.2 Methods

2.2.1 Likelihood Probability in Unascertained and Ascertained samples

FH includes phenotypes of parents or other relatives of probands. Suppose that genotypes and phenotypes of probands, and phenotypes of their relatives are available from a large cohort. The evidence for association can be assessed from two separate analyses: 1) association between probands' genotypes and their phenotypes; 2) the association between probands' genotypes and a set of relatives' phenotypes conditional on the probands' phenotypes. To derive the FHAT statistic, a weighted meta-analysis approach is used to combine the scores statistics from these two separate analyses. Both continuous and binary outcomes are handled by fitting appropriate models while adjusting for covariates. A linear model or linear mixed effect model is considered for continuous trait while a logistic model or logistic mixed model is considered for binary trait.

We assume that there are n probands with m observed variants included in the aggregation unit-based test. When we have FH on the relative of the probands, let Y_i^P denote the phenotype of the i^{th} proband; Y_i^R denote the phenotype of the relative of the i^{th} proband, respectively; G_i^P denote the genotypes of the i^{th} proband; X_i^P denote covariates for the i^{th} proband; X_i^R denote covariates of the relative of the i^{th} proband, such as age and ancestral principal components (PCs) that account for population structure. Different covariates can be adjusted for probands and relatives. In the application, we used the PCs from probands for the relatives. we assume that $X_i^P \perp Y_i^R$ conditional on X_i^R, and $X_i^R \perp Y_i^P$ conditional on X_i^P.

The likelihood for the unascertained data can be written as:

$$P(Y_i^P, Y_i^R \middle| G_i^P, X_i^P, X_i^R) = \frac{P(Y_i^P, Y_i^R, G_i^P, X_i^P, X_i^R)}{P(G_i^P, X_i^P, X_i^R)}$$

$$= \frac{P(Y_i^R \middle| G_i^P, Y_i^P, X_i^R, X_i^P) P(Y_i^P | G_i^P, X_i^P, X_i^R) P(G_i^P, X_i^P, X_i^R)}{P(G_i^P, X_i^P, X_i^R)}$$

$$= P(Y_i^R \middle| G_i^P, Y_i^P, X_i^R) P(Y_i^P | G_i^P, X_i^P). (1)$$

The first probability is the likelihood of family history given the proband information (and relative's covariate information) and the second probability is the prospective likelihood of the proband status given the proband's genotype and covariate information. For an ascertained sample, we can write the likelihood in an alternate form:

$$P(G_i^P, Y_i^R \middle| Y_i^P, X_i^P, X_i^R) = \frac{P(Y_i^P, Y_i^R, G_i^P, X_i^P, X_i^R)}{P(Y_i^P, X_i^P, X_i^R)} = \frac{P(Y_i^P, Y_i^R \middle| G_i^P, X_i^P, X_i^R) P(G_i^P, X_i^P, X_i^R)}{P(Y_i^P, X_i^P, X_i^R)}$$

$$= P(Y_i^R \middle| G_i^P, Y_i^P, X_i^R) \frac{P(Y_i^P | G_i^P, X_i^P) P(G_i^P, X_i^P)}{P(Y_i^P, X_i^P)}$$

$$= P(Y_i^R \middle| G_i^P, Y_i^P, X_i^R) \frac{P(Y_i^P, G_i^P, X_i^P)}{P(Y_i^P, X_i^P)}$$

$$= P(Y_i^R \middle| G_i^P, Y_i^P, X_i^R) P(G_i^P | Y_i^P, X_i^P),$$

where we also assume that $G_i^P \perp X_i^R$ conditional on X_i^P in addition to the assumptions for unascertained data. Again, the first probability is the likelihood of the family history given the proband information, which the second probability now refers to the standard likelihood for case-control data. Using this alternative likelihood formulation that takes ascertainment into consideration yield the same test statistics as (1), and hence our test is valid for case-control ascertained samples.

2.2.2 Incorporating FH from Both Parents for Single Variant Analysis

When FH of a single un-genotyped relative per proband is available, Ghosh et. al. [19] proposed a method to combine the FH of a single parent into the single variant analysis. We extend their method to handle FH on both parents per proband. Similar to (1), the likelihood of observing disease status of probands Y_i^P and disease status of mother Y_i^{R1} and father Y_i^{R2} conditional on probands' genotypes G_i^P is:

$$P(Y_i^P, Y_i^{R1}, Y_i^{R2} \middle| G_i^P, X_i^P, X_i^{R1}, X_i^{R2})$$
$$= P(Y_i^P \middle| G_i^P, X_i^P) P(Y_i^{R1} \middle| G_i^P, Y_i^P, X_i^{R1}) P(Y_i^{R2} \middle| G_i^P, Y_i^P, X_i^{R2}).$$

This equation allows to assess the combined analysis from fitting three regression models based on $P(Y^P|G^P)$, $P(Y^{R1}|G^P, Y^P)$ and $P(Y^{R2}|G^P, Y^P)$. Let to $\hat{\beta}^P$, $\hat{\beta}^{R1}$, and $\hat{\beta}^{R2}$ be the estimates of genetic effects from probands analysis and the other two relatives analyses, respectively. Due the underlying relationship between parents and offspring, the three test statistics from three regression models can be combined with the appropriate weights into a single test following the same framework in Ghosh et. al, $T = \frac{\frac{\hat{\beta}^P}{Var(\hat{\beta}^P)} + \frac{\hat{\beta}^{R1}}{2Var(\hat{\beta}^{R1})} + \frac{\hat{\beta}^{R2}}{2Var(\hat{\beta}^{R2})}}{\sqrt{\frac{1}{Var(\hat{\beta}^P)} + \frac{1}{4Var(\hat{\beta}^{R1})} + \frac{1}{4Var(\hat{\beta}^{R2})}}}$.

To illustrate the performance on power, we conducted the simulation analysis to evaluate the empirical power of the single variant analysis incorporating parental history from both parents and compared to the standard single variant analysis (without incorporating parental history). The analysis was restricted to common variants (MAF $\geq$ 1%) as the single variant analysis is underpowered for detecting rare variants, as shown in previous literature. [33,32] The empirical power was evaluated at a significance level of 0.0001 for n probands (n= 1000, 1500, 2500) with available parental history of disease status using 1000 simulation replicates. **Figure 2.1** shows that incorporating FH is a cost-effective way to improve power. The statistical power for the analysis of 1500 probands with parental FH incorporated is larger than the power of 2000 probands without FH information. The results from this analysis confirm the inclusion of FH will increase statistical power compared to the standard analysis in the context of single common variants.

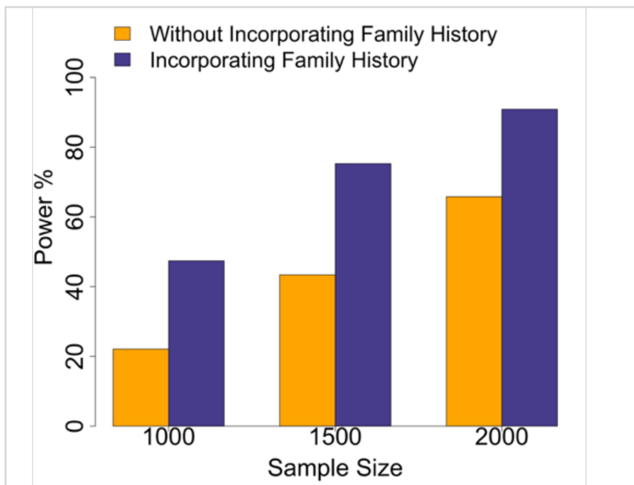

Figure 2.1 Statistical power for single variant analysis

2.2.3 Family History Aggregation unit-based Tests (FHAT)

We propose a novel approach called FHAT to incorporate FH information in the aggregation unit-based tests using the variance component test framework. Based on $P(Y_i^P | G_i^P)$, we first assess the association between probands' genotypes and their disease status using

$$g(E(Y_i^P | G_i^P, X_i^P)) = X_i^P \alpha_P + G_i^P \beta_P, \quad (2)$$

where $g(\cdot)$ is the link function, α_P is a vector of regression coefficients for covariate effects, β_P is a vector of regression coefficients for the observed genotypes in probands. The model for relatives based on $P(Y_i^R \middle| G_i^P, Y_i^P)$ is specified as

$$g(E(Y_i^R \mid G_i^P, Y_i^P, X_i^R)) = X_i^R \alpha_R + G_i^P \beta_R + Y_i^P \lambda_R, \quad (3)$$

where λ_R is scalar of regression coefficients for probands' phenotypes for the relatives' model; α_R is vector of regression coefficients for relatives' covariates; β_R is vector of regression coefficients for m observed variants in probands. This relatives' model (3) can analyze FH from unrelated relatives, i.e. single relative per probands or FH from both parents since mothers and fathers are conditional independent. The two underlying association estimators, $(\hat{\beta}_P, \hat{\beta}_R)$, have the following relationship [55]

$$\hat{\beta}_R \approx 2\Omega\hat{\beta}_P,$$

where Ω is the kinship coefficient between probands and their relatives and $\Omega = \frac{1}{4}$ for first-degree relatives such as parents.

Conventional aggregation unit-based methods evaluate the association between a set of variants and phenotype. One such aggregation unit-based method is called the SKAT [59], which is a variance component score test in a mixed model framework for rare variant associations. SKAT was derived under the assumption that the genetic effect β_j for variant j follows an arbitrary distribution with mean zero and a variance of $w_j^2\tau$, where w_j is a pre-specified weight for variant j and τ is a variance component. In SKAT, testing whether the set of variants is associated with a phenotype, $H_0\colon \beta_P = 0$, corresponds to testing $H_0\colon \tau =$

0. Specifically, the weighted SKAT statistic based on the probands' model (2) is defined as

$$Q_{SKAT} = \frac{(Y^P - \hat{\mu}_P)^T G^P W W G^{P^T} (Y^P - \hat{\mu}_P)}{\hat{\phi}_P^2},$$

where $W = diag(w_1, w_2, \ldots, w_m)$ is a pre-specified weight matrix for m variants; G^P is a $n \times m$ genotype matrix with $(i, j)_{th}$ element corresponding to the additively coded genotype for variant j of proband i; $\hat{\mu}_P$ is the estimated mean of Y^Pusing the null model with only covariates; $\hat{\phi}_P$ is the estimate of dispersion parameter under H_0 and $\hat{\phi}_P = 1$ for binary trait. The SKAT statistic can be obtained similarly to evaluate whether genetic variants are associated with disease status using the relatives' phenotypes to replace the probands' phenotypes based on relatives' model (3). In both probands and relatives analyses, the pre-specified weights are typically a function of the minor allele frequency. For example, one can use Wu's weights [59] $w_j = Beta\ (MAF_j; 1, 25)$ to up-weight the effect of rarer variants.

We propose to combine the score statistics from the two association models for probands and their relatives using a weighted meta-analysis. Meta-analysis is often used in genetic association analysis to increase the power by combining results from multiple studies. Methods to meta-analyze SKAT results have been developed. [36] Meta-analysis of rare variant association tests proposed are based on the study-specific summary statistics, that is, score statistics for each variant and linkage disequilibrium estimates in a region. Because of the genetic relationship between probands and their relatives, we down-weight the scores for relatives by 2Ω based on their relationship when combining the score statistics in a meta-analysis by assuming the homogeneous genetic effects among probands and their relatives. Specifically, because relative k of each proband may or may not have phenotype data available, we use Y^{R_k} to denote the collective phenotype vector for relative k of all probands (e.g., all mothers), including missing values, with kinship coefficient Ω_k. The diagonal matrix $D(R_k)$ indicates whether corresponding element in Y^{R_k} for each proband is missing (denoted by 0) or not (denoted by 1). Therefore, relatives with missing phenotype data do not contribute to the test statistic. We fit a single relative model jointly using all relatives' phenotypes and covariates conditional on their probands' phenotypes to get $\hat{\mu}_{R_k}$, the estimated mean vector of Y^{R_k} for relative k of all probands, as well as the dispersion parameter estimate $\hat{\phi}_R$ under the null hypothesis of no genetic effects. We assume that all relatives are independent in the model. The general form of FHAT statistics that incorporates FH from relatives is

$$Q_{FHAT} = \left[\frac{(Y^P-\hat{\mu}_P)^T}{\hat{\phi}_P} + \sum_k \frac{2\Omega_k D(R_k)(Y^{R_k}-\hat{\mu}_{R_k})^T}{\hat{\phi}_R}\right] G^P W\, W G^{P^T} \left[\frac{(Y^P-\hat{\mu}_P)}{\hat{\phi}_P} + \sum_k \frac{2\Omega_k D(R_k)(Y^{R_k}-\hat{\mu}_{R_k})}{\hat{\phi}_R}\right] = \sum_{j=1}^{m}\left(w_j^2 S_{Pj}^2 + \sum_k 4{\Omega_k}^2 w_j^2 S_{R_{k_j}}^2\right) \quad (4)\,,$$

where S_{Pj} is the score statistic for probands for variant j, and $S_{R_{k_j}}$ is the score statistic for relative k of all probands for variant j. Specifically,

$$S_{Pj} = \sum_{i=1}^{n} g_{ij}(Y_i^P - \hat{\mu}_{Pi})/\hat{\phi}_P,$$

$$S_{R_{k_j}} = \sum_{i=1}^{n} d_i(R_k) g_{ij}(Y_i^{R_k} - \hat{\mu}_{R_{k_i}})/\hat{\phi}_R,$$

where $d_i(R_k)$ is the ith diagonal element from $D(R_k)$ indicating relative k for proband i is missing (denoted by 0) or not (denoted by 1), g_{ij} is the genotype (coded as 0, 1, or 2) for the j_{th} variant of i_{th} proband, $\hat{\mu}_{Pi}$ and $\hat{\mu}_{R_{k_i}}$ are the estimated means of Y_i^P and $Y_i^{R_k}$ for of i_{th} proband and the relative of i_{th} proband the under the null model, respectively, and $\hat{\phi}_P$ and $\hat{\phi}_R$ are the dispersion parameter estimates for probands and relatives, respectively.

Let $V_P = (w_1 S_{P1}, w_2 S_{P2}, \dots, w_m S_{Pm},)^T$ and $V_R = \left(\sum_k 2\Omega_{\mathrm{k}} w_1 S_{R_{k_1}}, \sum_k 2\Omega_{\mathrm{k}} w_2 S_{R_{k_2}}, \dots, \sum_k 2\Omega_{\mathrm{k}} w_m S_{R_{k_m}},\right)^T$,then $V = V_P + V_R$ follows a multivariate normal distribution of mean 0 and covariance matrix of

$$\Psi = W\, G^{P^T}\left(\hat{P} + \sum_k 4\Omega_{\mathrm{k}}^2 D(R_k)\hat{P}_{R_k}D(R_k)\right)G^P W\,,$$

where $\hat{P} = \hat{\Sigma}^{-1}{}_P - \hat{\Sigma}^{-1}{}_P X_P\left(X_P^T\hat{\Sigma}^{-1}{}_P\, X_P\right)^{-1}X_P^T\hat{\Sigma}^{-1}{}_P$, $\hat{P}_{R_k} = \hat{\Sigma}^{-1}{}_{R_k} - \hat{\Sigma}^{-1}{}_{R_k}X_{R_k}\left(X_{R_k}^T\hat{\Sigma}^{-1}{}_{R_k}\, X_{R_k}\right)^{-1} X_{R_k}^T\hat{\Sigma}^{-1}{}_{R_k}$ are the projection matrices in probands and relatives k, respectively, and $\hat{\Sigma}_P = \hat{\phi}_P\boldsymbol{I}$ and $\hat{\Sigma}_{R_k} = \hat{\phi}_R\boldsymbol{I}$ for continuous traits and $\hat{\Sigma}_P = diag\{1/\hat{\mu}_{Pi}(1-\hat{\mu}_{Pi})\}$ and $\hat{\Sigma}_{R_k} = diag\{1/\hat{\mu}_{R_{k_i}}(1-\hat{\mu}_{R_{k_i}})\}$ for the binary traits. Therefore,

$$Q_{FHAT} = V^T V \sim \sum_{j=1}^{m}\lambda_j\chi_{1,j}^2,$$

which follows a weighted sum of chi-square distribution with 1 degree of freedom. The weights λ_j can be estimated from the eigenvalues of Ψ. The p-value can be estimated by the Davies' method [11]. The general form (4) can be reduced to

$$Q_{FHAT} = \left[(Y^P - \hat{\mu}_P)^T + \frac{D(R_m)(Y^{R_m} - \hat{\mu}_{R_m})^T}{2} + \frac{D(R_f)(Y^{R_f} - \hat{\mu}_{R_f})^T}{2}\right] G^P W \, W G^{P^T} \left[(Y^P - \hat{\mu}_P) + \frac{D(R_m)(Y^{R_m} - \hat{\mu}_{R_m})^T}{2} + \frac{D(R_f)(Y^{R_f} - \hat{\mu}_{R_f})^T}{2}\right] \quad (5)$$

for incorporating FH from both parents (with mothers denoted by m and fathers denoted by f) when using logistic models for binary trait with the estimates of dispersion parameters fixed to 1 (i.e., $\hat{\phi}_P = \hat{\phi}_R = 1$), and the kinship coefficients ($\Omega_{\mathrm{m}}, \Omega_{\mathrm{f}}$) fixed to $\frac{1}{4}$.

2.2.4 Optimal FHAT (FHAT-O)

Using the same weighted-meta analysis framework adopted in FHAT, we developed a FHAT-O statistic based on the optimal unified test SKAT-O [37]. SKAT outperforms Burden tests [38,42,47] when the effects of variants are in different directions. However, when the proportion of causal variants is high and variants have positive effects, SKAT suffers from a loss of power compared to Burden tests. Since SKAT-O combines the feature of SKAT and Burden tests, the power is robust in the presence of both different and same directions of causal variant effects. To test the hypothesis $H_0 : \beta_P = 0$, the Burden score statistic derived using model (2) is

$$Q_{Burden} = \left[\frac{1}{\hat{\phi}_P} \sum_{i=1}^{n} (Y_i^P - \hat{\mu}_{Pi}) \left(\sum_{j=1}^{m} w_j g_{ij}\right)\right]^2.$$

FHAT-Burden is proposed based on Burden test using the FHAT framework to incorporate FH, which can be represented as a weighted sum of the weighted score statistics in probands, and relatives based on their relationships,

$$Q_{FHAT-Burden} = \left[\frac{\sum_{i=1}^{n}(Y_i^P - \hat{\mu}_{Pi})\left(\sum_{j=1}^{m} w_j g_{ij}\right)}{\hat{\phi}_P}\right.$$

$$\left. + \sum_k \frac{2\Omega_k \sum_{i=1}^{n} d_i(R_k)\left(Y_i^{R_k} - \hat{\mu}_{R_{k_i}}\right)\left(\sum_{j=1}^{m} w_j g_{ij}\right)}{\hat{\phi}_R}\right]^2,$$

where $d_i(R_k)$ is the ith diagonal element from $D(R_k)$, indicating relative k for proband i is missing (denoted by 0) or not (denoted by 1). A unified test defining as the weighted average of FHAT and FHAT-Burden is

$$Q_\rho = (1-\rho)Q_{FHAT} + \rho Q_{FHAT-Burden}$$

$$= (1-\rho)\sum_{j=1}^{m}(w_j^2 S_{Pj}^2 + \sum_k 4\Omega_k{}^2 w_j^2 S_{R_{k_j}}^2) + \rho\sum_{j=1}^{m}(w_j\, S_{Pj} + \sum_k 2\Omega_k\, w_j\, S_{R_{k_j}})^2$$

$$= (1-\rho)V^T V + V^T\rho\mathbf{1}\mathbf{1}^T V \sim \sum_j^m \lambda_{\rho j}\chi_{1,j}^2,$$

which asymptotically follows the distribution of a mixture-squire distribution under the null. In the summation, $\lambda_{\rho j}$s are the eigenvalues of $L_\rho^T \Psi L_\rho$ where $L_\rho$is the matrix satisfying $L_\rho L_\rho{}^T = (1-\rho)\boldsymbol{I} + \rho\mathbf{1_m}\mathbf{1_m}{}^T$, where $\mathbf{1_m}$ is the vector of length m with all elements 1. Let P_ρ denote the p-value of Q_ρ for a given ρ, the test statistic for FHAT-O that combines the features of FHAT and FHAT-Burden is determined as

$$Q_{FHAT-O} = \min_{0\le\rho\le 1} P_\rho, \quad (6)$$

where P_ρ is the p-value estimated for each given ρ. Defining a grid set for ρ as $0 < \rho_1 < \rho_2 < \cdots < \rho_L$, then $Q_{FHAT-O=}$min $(P_{\rho_1}, P_{\rho_2}, \ldots, P_{\rho_L})$. It can be shown that $Q_\rho(\rho)$ is the mixture of two independent random variables: one follows a chi-square distribution with df = 1, the other one is asymptotically approximated to the mixture of chi-square distribution with adjustment. We can use the approach proposed by Lee et al. [37] to obtain the optimal ρ. When ρ = 1, Q_ρ reduces to FHAT-Burden, and when ρ = 0, Q_ρ is equivalent to FHAT.

2.2.5 Liability Threshold Model of Case–Control Status and Family History in Rare Variant Analysis

A method that utilizes posterior mean generic liabilities under the liability threshold model of case-control status and FH (LT-FH) has been proposed to increase association power by combining case-control status and available FH, where they demonstrated 63% and 36% increases in power compared to GWAS and genome-wide association by proxy (GWAX) [39], respectively. [26] To make a comparison between LT-FH approach to our methods (FHAT and FHAT-O) for rare variant analysis, we first calculate posterior mean genetic liabilities $Y_{LT\text{-}FH}$ using LT-FH conditional on both probands' disease status and FH from relatives. Then we incorporate the LT-FH phenotype $Y_{LT\text{-}FH}$ as a continuous outcome in SKAT and SKAT-O to test the association between aggregated groups of rare variants and the phenotype of interest.

2.3 Simulation Studies

Simulations were performed to evaluate the FHAT and FHAT-O statistics in terms of empirical type I error and statistical power. We generated 10,000 haplotypes for a 4kb region on chromosome 19 using HapGen2 software [54]. The data from 1000 genomes project was used as the reference panel to simulate haplotypes. In all simulations, we focused on binary traits because they are more often collected through questionnaire in relatives and we focused on rare variants with MAF < 1%. We simulated the probands with both genotypes and phenotypes, and with available FH data from both of their parents. We used LT-FH phenotype in SKAT (SKAT-LTFH) and SKAT-O (SKATO-LTFH) and compared the results to FHAT and FHAT-O, and they were all calculated by combining the FH from relatives (i.e. mothers and fathers) into the analysis. The standard methods (SKAT, SKAT-O, Burden tests and ACAT-V) only used proband data. Because mothers and fathers were simulated as independent samples, they were analyzed using a single relatives model (3) and then FHAT and FHAT-O statistics were calculated using (5) and (6). We used Davies' method for evaluating p-values of a weighted sum of chi-square distributions. The type I error and power of FHAT and FHAT-O were compared to SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, Burden tests and ACAT-V. Note that ACAT-V is an aggregation unit-based test for rare variants associations combining variant-level p-values using aggregated Cauchy association test (ACAT). [40]

2.3.1 Type I Error

2.3.1.1 Type I Error Simulation Design

Several simulation analyses were conducted under the null hypothesis of no genetic associations. The type I error was estimated using different pre-specified disease prevalence and calculated at various alpha levels. A total of 10,000 simulated haplotypes were sampled to generate 5000 probands, each assigned two haplotypes at random. Then, a null phenotype not associated with the genotypes, Y^P, for 5000 probands and their mothers Y^M and fathers Y^F was simulated for each replicate using the following model,

$$\begin{pmatrix} Y^P \\ Y^M \\ Y^F \end{pmatrix} = 0.015 \begin{pmatrix} age^P \\ age^M \\ age^F \end{pmatrix} + 0.25 \begin{pmatrix} sex^P \\ sex^M \\ sex^F \end{pmatrix} + \varepsilon, \qquad \varepsilon \sim MVN(0, \Sigma),$$

where $age^P, age^M$and age^F are vectors of continuous variable randomly selected from ages of probands, mothers and fathers in UK Biobank data, respectively; sex^P is a vector of binary variable generated from a Bernoulli distribution with probability for female = 56% in probands; sex^M and sex^F are fixed as female and male, respectively; ε is the error term following a multivariate normal distribution with mean of zero and covariance matrix of Σ,

$$\Sigma = \delta_g^2 \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix} + \delta_e^2 I_{3\times3},$$

and we set $\delta_g^2 = \delta_e^2 = 0.5$. The binary phenotype was generated using different cut-offs from the simulated continuous phenotype data to have pre-specified prevalence p in probands. We fixed p , the prevalence in probands, at ~ 20% and 50%, with increased prevalence in mothers (~ 35% and 69%) and fathers (~ 28% and 62%), to mimic real AD/dementia data, with more prevalent disease in women and older relatives. We used $Beta\ (MAF_j; 1, 25)$ weights in FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, and SKAT-O. The comparable weights of $w_{j,ACAT-V} = w_{j,SKAT} \times \sqrt{MAF_j(1 - MAF_j)}$ were used in ACAT-V. [40] The type I error rates were evaluated by the proportion of p-values less than or equal to the alpha levels of 2.5×10^{-2}, 2.5×10^{-3}, 2.5×10^{-4}, and 2.5×10^{-5}.

2.3.1.2 Type I Error Simulation Results

A total of 2 million simulation replicates were first generated to evaluate type I error at various alpha levels for FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, Burden test and ACAT-V using 5000 probands with available parental history (**Table 2.1**). When the disease prevalence is low, SKAT and SKAT-O have inflated type I error for prevalence = 20% and alpha =2.5×10^{-4} and 2.5×10^{-5}, while the type I error is controlled better in FHAT, FHAT-O, SKAT-LTFH, and SKATO-LTFH when combining additional cases in relatives. When the prevalence is set to 50%, a slightly deflated type I error was observed in FHAT, SKAT-LTFH, and SKAT in some scenarios. The conservativeness of SKAT when the prevalence is 50% was also observed in prior publications. [37, 59]

The type I error evaluation results for other disease prevalence and alpha levels (including exome-wide significance) can be found in **Table A.1** in Appendix A.1. We demonstrated that both FHAT and FHAT-O have reasonable type I error at the exome-wide significance (alpha= 2.5×10^{-6}) and other disease prevalence, and when the prevalence is low, FHAT and FHAT-O have lower inflation while SKAT and SKAT-O suffer from substantial inflated type I error.

Table 2.1 Type I error rates of FHAT, SKAT- LTFH, SKAT, FHAT-O, SKAT- LTFH, SKAT-O, Burden, and ACAT-V

Alpha	FHAT	SKAT-LTFH	SKAT	FHAT-O	SKATO-LTFH	SKAT-O	Burden	ACAT-V
Prevalence = 20%								
2.5×10^{-2}	0.96	0.98	1.00	0.99	1.02	1.02	1.00	1.15
2.5×10^{-3}	0.93	0.98	1.06	1.07	1.11	1.20	1.01	1.14
2.5×10^{-4}	0.91	1.00	1.27	1.17	1.27	1.56	1.03	1.16
2.5×10^{-5}	1.14	1.14	1.92	1.66	1.82	2.38	0.98	1.24
Prevalence = 50%								
2.5×10^{-2}	0.92	0.96	0.94	0.98	1.00	0.99	1.00	1.15
2.5×10^{-3}	0.89	0.91	0.88	1.04	1.06	1.03	1.00	1.07
2.5×10^{-4}	0.85	0.82	0.81	1.01	1.09	1.05	0.92	0.87
2.5×10^{-5}	0.62	0.58	0.72	1.00	1.02	1.08	0.78	0.68

The number in each cell represents the ratio of type I error and expected significance level (column 'Alpha'). Type I error was evaluated from the proportion of p-values less than or equal to corresponding alpha level= 2.5 $\times 10^{-2}$, 2.5×10^{-3}, 2.5×10^{-4} and 2.5×10^{-5} using 2 million simulation replicates. FHAT, SKAT-LTFH, SKAT, FHAT-O, SKATO-LTFH, SKAT-O and Burden test all used the same Wu weights with beta (MAF_j; 1, 25). ACAT-V used the weight of $w_{j,ACAT-V} = w_{j,SKAT} \times \sqrt{MAF_j\,(1 - MAF_j)}$ to make results comparable. The analyses were restricted to rare variants with MAF< 1%. The LTFH phenotype was computed using LT-FH software v2 and then used as the continuous outcome in SKAT and SKAT-O to obtain SKAT-LTFH and SKATO-LTFH.

2.3.2 Power

2.3.2.1 Power Simulation Design

The power was assessed under various scenarios. We randomly assigned two haplotypes from 10,000 simulated haplotypes to each parent, and each parent then randomly passed one of the two haplotypes without recombination to the probands. We considered a model with an interaction between age and variants to simulate a continuous phenotype

$$\begin{pmatrix} Y^P \\ Y^M \\ Y^F \end{pmatrix} = 0.015 \begin{pmatrix} age^P \\ age^M \\ age^F \end{pmatrix} + 0.25 \begin{pmatrix} sex^P \\ sex^M \\ sex^F \end{pmatrix} + 0.015 \begin{pmatrix} {G_{causal}}^P \\ {G_{causal}}^M \\ {G_{causal}}^F \end{pmatrix} \gamma \begin{pmatrix} age^P \\ age^M \\ age^F \end{pmatrix} + \varepsilon,$$

where age, sex, and ε are defined in the model used for type I error simulations; G_{causal}^P, G_{causal}^M and G_{causal}^F are the genotype matrices of causal variants for probands, mothers and fathers in the true model, respectively, but we assumed that G_{causal}^M and G_{causal}^F were missing when we calculated FHAT, FHAT-O and SKAT-LTFH and SKATO-LTFH; γ is a vector of effect sizes for the causal variants. The elements in vector γ are specified as following [6],

$$\gamma_j = \sqrt{\frac{c}{2MAF_j(1 - MAF_j)}},$$

where MAF_j is the MAF of causal variant j, c is a pre-specified constant and defined as

$$c = \frac{R^2}{V^T DV},$$

where R^2 is the proportion of variance explained by causal variants, we set to 2% in our simulation when all causal variants have same effect directions, and 5% when half of the causal variants have positive effects and half of the causal variants have negative effects on the liability scale for binary trait; D consists of the LD correlation matrix between variants; V is a vector corresponding to the effect directions of causal variants. We simulated data by varying the number of total variants testing in a region, and the proportion of causal variants. The same strategy was used to generate binary phenotypes described in the type I error simulations: the binary phenotype was generated using different cut-offs for simulated continuous phenotype data.

The comparable weights were used to calculate statistics for FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, Burden tests and ACAT-V. We considered scenarios where the number of variants analyzed in a region was 20, 40, and 80, the disease prevalence $p = 20\%$ and 50%, and the proportion of causal variants is 10%, 20%, 50%, 80% and 100%. The power was calculated as the proportion of p-values less than or equal to the alpha level = 2.5×10^{-5} and exome-wide significance level = 2.5×10^{-6} for testing 20,000 genes.

2.3.2.2 Power Simulation Results

Figure 2.2 summarizes the power simulation results of FHAT, SKAT-LTFH, SKAT, FHAT-O, SKATO-LTFH, SKAT-O, Burden tests and ACAT-V for disease prevalence = 20% and alpha= 2.5×10^{-5}. (See Appendix A.2 for additional power results). The causal variants in a region were set to have positive effects, or half of the causal variants have positive effects and half of the causal variants have negative effects. In all scenarios, similar patterns are shown in **Figure 2.2** and **Figure 2.3**. Our main findings included: 1) FHAT and FHAT-O are slightly more powerful than SKAT-LTFH and SKATO-LTFH, respectively, under many scenarios when the variants have larger effects on the disease among older people; 2) FHAT and FHAT-O have greatly improved power compared to standard method that do not incorporate FH in most scenarios except for the scenario when the proportion of causal variants is 10% and half of the causal variants have positive effects and half of the causal variants have negative effects. However, ACAT-V has substantial power loss in many other scenarios; 3) FHAT suffers from a loss of power when the proportion of causal variants is high and the causal variants have effects in the same directions. In contrast, FHAT-O outperforms FHAT in those scenarios, and remains powerful regardless of the directions of genetics effects or number of causal variants.

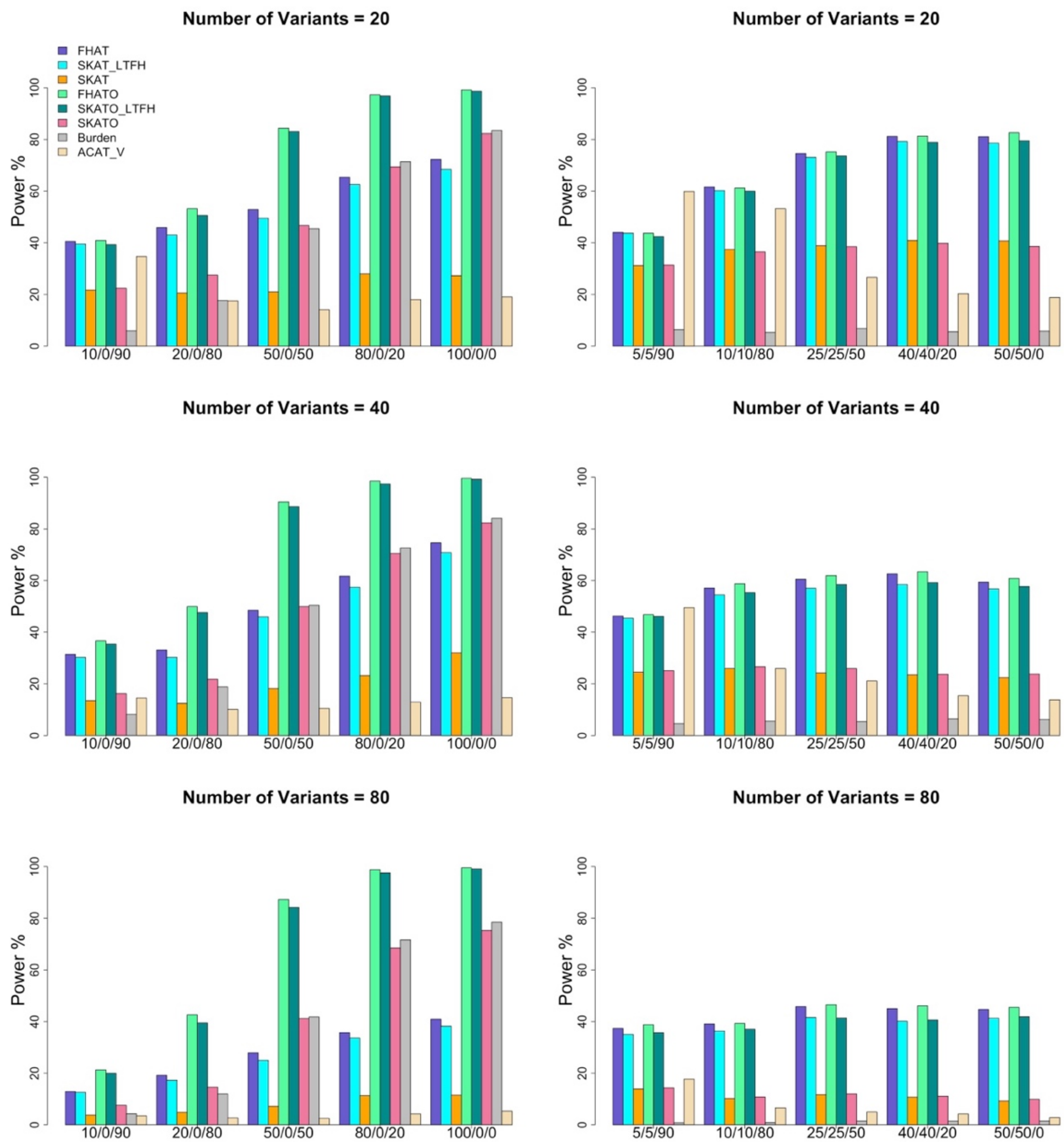

Figure 2.2 Empirical power of FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, Burden test and ACAT-V estimated at alpha= 2.5 x 10^{-5} for prevalence = 20%

In each plot, the x axis in the format of +/-/0 indicates the proportion of variants with positive, negative and no effects. Each bar shows the empirical power evaluated as the proportion of p-values less than or equal to alpha= 2.5 x 10^{-5}. FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, and Burden test all used the same Wu weights with beta (MAF_j; 1, 25). ACAT-V used the weight of $w_{j,ACAT-V} = w_{j,SKAT} \times \sqrt{MAF_j(1 - MAF_j)}$ to make results comparable. The analyses were restricted to rare variants with MAF< 1%.

2.3.3 Computational Cost

FHAT and FHAT-O have lower computational cost compared to SKAT-LTFH and SKATO-LTFH. **Table 2.2** summarizes computation time (in minutes) for all methods for analyzing 1000 regions that contain 30 variants. The computation time of FHAT, FHAT-O, SKAT, SKAT-O, Burden tests and ACAT-V depends on sample size and region size, whereas the running time for SKAT-LTFH and SKATO-LTFH (conducting using the LT-FH software v2 [26]) depends on the number of configurations of probands' disease status and FH.

Table 2.2 Computational time for testing 1000 regions

Sample Size	FHAT	FHAT-O	SKAT	SKAT-O	Burden	ACAT-V	SKAT-LTFH	SKATO-LTFH
200	0.09	0.25	0.12	1.38	0.13	0.04	540.33	544.83
500	0.14	0.33	0.14	1.57	0.16	0.06	536.42	543.09
1000	0.26	0.43	0.24	1.69	0.23	0.09	534.54	541.84
2000	0.53	1.25	0.42	2.56	0.39	0.14	566.45	568.20
5000	1.19	1.89	0.90	5.40	0.81	0.29	551.78	553.74

Each cell summarizes the time (in minutes) that is required to preforming the tests on 100 regions using the methods of FHAT, SKAT- LTFH, SKAT, FHAT-O, SKAT- LTFH, SKAT-O, Burden, and ACAT-V. The regions contain 30 variants.

2.4 Application to the UK Biobank

2.4.1 Analysis of Whole Exome Sequencing Data

We applied FHAT and FHAT-O to analyze all cause dementia and hypertension in the UK Biobank data. Rare variant (with MAF < 1%) gene-based analyses were conducted. The UK Biobank is a large prospective cohort study with information on clinical traits, covariates, and genome-wide genotype data for over 500,000 individuals with age at assessment between 37-73 years at baseline (2006 to 2010). The UK Biobank has released the first and second tranches of exome sequencing data for ~200,000 samples who were sequenced by the UK Biobank Exome Sequencing Consortium with the GRCh38 reference.

FH of all cause dementia and hypertension was collected from questionnaires. The information of FH was collected from participants by answering the questions of "Has/did your father/mother/brothers/sisters ever suffer from any of the following diseases". We applied the methods by incorporating the FH of all cause dementia (including dementia and AD) from both parents in the whole exome-wide analysis. We adjusted the all cause dementia analysis for age and sex, and adjusted the hypertension analysis for age, age squared (age^2), sex, and body mass index (BMI). The BMI of probands was used as a proxy for BMI for the parents. To account for population structure, we additionally adjusted ancestral PCs in the analysis. The PCs from probands were used in parental analysis. We first tested the top 20 PCs in probands and parental analysis, and we meta-analyzed the results to determine which PCs to include in all cause dementia and hypertension analyses (Appendix A.3, Supplementary **Table A.2**). We included the top 5 PCs and any additional PCs reaching statistical significance ($P< \frac{0.05}{20} = 2.5\times10^{-3}$) in the models: the PC1-PC5 and PC11 were included in the all cause dementia analyses, while PC1-PC5, PC8, and PC14 were included in models for hypertension. We did the variant-level quality control (QC) by removing any variants with missing rate > 5%, and then we imputed the missing genotypes using the mean genotype values, as is standard in the SKAT method. The following QC procedures were implemented to select samples: samples with high heterozygosity and high missing rates, sex aneuploidy, and mismatches reported sex with affymetrix-determined sex were removed from our analysis. We selected frameshift, splice accepter, splice donor, stop gained, stop lost, and missense variants determined by Ensemble Variant Effect Predictor (VEP) annotation for inclusion in our analyses. We used the genotypes generated from Functionally Equivalent (FE) pipeline and we omitted regions that are affected by non-alt-aware mapping in the UK Biobank exome sequencing data.

2.4.2 Results

There are 129,670 unrelated white individuals (with ethnic background of White, British, Irish, and Any other white background) who passed all QC filters and have exome sequencing data, phenotype, and available parental disease status. The unrelated individuals are identified by removing samples who have multiple 1st, 2nd, and 3rd degree relatives, randomly omitting one sample from each relative pair. The age at the first assessment visit for probands is between 38 and 72 with the mothers of probands being between 60 and 105, and the fathers of probands being between 60 and 102. There are 27 dementia cases (p=0.02%) and 32,773 hypertension cases (p=25.3%) among probands. While mothers and fathers of probands have similar hypertension prevalence (37,145 hypertension cases in mothers, p=28.6%; 26,063 hypertension cases in fathers, p= 20.1%), more dementia cases are observed in the parents (10,654 dementia cases in mothers, p=8.2%; 5,720 dementia cases in fathers, p= 4.4%) compared to probands.

We applied our methods to exome sequencing data in the UK Biobank. We first evaluated the associations between all cause dementia and hypertension with known regions previously implicated with AD/dementia risk and hypertension. We performed the analysis for all unrelated white individuals using FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH and other conventional aggregation unit-based tests (SKAT, SKAT-O, Burden tests, and ACAT-V), see results in **Table 2.3**. The samples involved in the analyses varied because of missing values in the covariates used for adjustment in the models. FHAT, SKAT-LTFH, FHAT-O and SKATO-LTFH had improved significance after incorporating parental phenotype information compared to p-values calculated using other conventional tests for majority of genes. SKAT, SKAT-O and ACAT-V had almost no power to detect some associations for all cause dementia due to low prevalence in probands. The results show that *BCL3* (P= 6.8×10^{-5} in FHAT, P= 2.5×10^{-5} in SKAT-LTFH, P= 5.9×10^{-5} in FHAT-O, P= 1.8×10^{-5} in SKATO-LTFH) and *TOMM4* (P= 3.0×10^{-4} in FHAT, P= 5.8×10^{-4} in SKAT-LTFH, P= 3.8×10^{-4} in FHAT-O, P= 7.7×10^{-4} in SKATO-LTFH) were significantly associated with all cause dementia status at a significance level of 6.3×10^{-3} for testing 8 genes. At the same significance level, *DBH* (P= 1.3×10^{-3} in FHAT, P= 2.0×10^{-3} in SKAT-LTFH, P= 2.6×10^{-3} in FHAT-O, P= 3.3×10^{-3} in SKATO-LTFH) was associated with hypertension and had improved significance compared to the results from conventional methods. Although the tests that incorporate FH demonstrated an improved significance for all 8 AD/dementia genes we tested, some p-values for hypertension genes were less significant. This may be due to the fact that the prevalence for hypertension in probands was similar to that in parents, and the associations were diluted by the potential noises that were added when combining the FH from parents.

Table 2.3 Association analysis results for genes previously implicated in all cause dementia and hypertension susceptibility

Gene	FHAT	SKAT-LTFH	SKAT	FHAT-O	SKAT-O-LTFH	SKAT-O	Burden	ACAT-V	#variants	cumMAC	cumMAC in cases
All cause dementia (N=129,670)											
BCL3	6.8×10^{-5}	2.5×10^{-5}	0.02	5.9×10^{-5}	1.8×10^{-5}	0.029	0.11	0.36	65	1157	1
TOMM40	3.0×10^{-4}	5.8×10^{-4}	1	3.8×10^{-4}	7.7×10^{-4}	0.85	0.68	0.05	39	809	0
APOE	0.02	0.02	1	0.03	0.04	0.83	0.60	0.06	75	1372	0
PILRA	0.17	0.15	0.93	0.27	0.25	1.0	0.99	0.88	48	4406	1
BIN1	0.61	0.70	1	0.77	0.88	0.89	0.64	0.75	80	975	0
CR1	0.33	0.26	1	0.51	0.42	0.38	0.21	0.03	310	8500	0
CLU	0.44	0.43	1	0.59	0.59	0.66	0.45	0.79	75	2651	0
MAF	0.50	0.44	1	0.60	0.43	0.88	0.65	0.43	64	930	0
Hypertension (N=129,206)											
DBH	1.3×10^{-3}	2.0×10^{-3}	3.8×10^{-3}	2.6×10^{-3}	3.3×10^{-3}	1.6×10^{-3}	3.0×10^{-3}	0.05	166	7708	1851
SVEP1	0.051	0.09	0.066	0.068	0.10	0.091	0.082	0.36	485	19123	4707
NPR1	0.069	0.06	0.042	0.026	0.03	6.9×10^{-3}	4.0×10^{-3}	0.06	147	2739	749
REN	0.069	0.12	0.22	0.13	0.22	0.38	0.83	0.45	68	586	145
NPPA	0.21	0.27	0.40	0.28	0.40	0.48	0.33	0.03	31	1200	309
CHDH	0.24	0.25	0.54	0.29	0.33	0.33	0.20	0.75	120	2346	556
NF1	0.28	0.43	0.27	0.44	0.63	0.43	0.51	0.43	325	7826	1944
AGPS	0.36	0.33	0.38	0.54	0.50	0.56	0.98	0.88	172	5748	1449
PABPC4	0.93	0.98	0.91	0.42	0.49	0.11	0.060	0.79	60	283	61

The all cause dementia model adjusted age, sex, and PC1-5, and PC11 as covariates. The hypertension model adjusted age, age squared, sex, BMI, PC1-5, PC8 and PC14 as the covariates. Wu weights with beta (MAF_j; 1, 25). were used. The significance threshold is $\frac{0.05}{8} = 6.3 \times 10^{-3}$. cumMAC is the cumulative minor allele counts in probands for the gene we tested. #variants is the total number of variants tested in the gene.

A comprehensive exome-wide analysis was then conducted. A total of ~18K genes with two or more rare genetic variants meeting our filtering criteria were included. We used models including the same covariates for all cause dementia and hypertension as we did in the known gene analyses. In the analysis of all cause dementia (**Table 2.4, Figure 2.3**), the gene *TREM2* [23] (P= 4.1×10^{-9}) with known effects on AD/dementia and late onset AD achieved a strict exome-wide significance (P< 2.8×10^{-6}) using FHAT-O and it was also detected by FHAT (P= 5.2×10^{-6}) with a suggestive exome-wide significance (P< 5.6×10^{-5}). One known AD/dementia gene, *PVR* [27] (P= 1.2×10^{-5} in FHAT and P= 1.8×10^{-5} in FHAT-O) was identified with both FHAT and FHAT-O analysis, and *ABCA7* [25] (P= 4.1×10^{-5}) with known effects on AD/dementia was identified by FHAT-O. Moreover, three novel genes were found to be significantly associated with all cause dementia using FHAT and FHAT-O (*EFCAB3* with P= 4.0×10^{-5} in FHAT and P= 4.2×10^{-5} in FHAT-O, *EMSY* with P= 4.4×10^{-5} in FHAT and P= 2.7×10^{-5} in FHAT-O, and *KLC3* with P= 1.4×10^{-5} in FHAT-O). Because we observed highly inflated results (**Figure 2.3**) from hypertension analysis due to the correlation among parents' phenotypes, we corrected the analysis by adjusting for the spouse's hypertension status in the parents' model. For the adjusted hypertension analysis (**Table 2.4, Figure 2.3**), FHAT identified *GATA5* (P= 4.1×10^{-5}), and FHAT-O identified *FGD5* (P= 4.3×10^{-5}) and *DDN* (P= 4.2×10^{-5}) at a suggestive significance level. Those genes detected by our methods have previously been reported to be associated with hypertension and hypertension-related trait. [15,22,46]

Table 2.4 Whole exome-wide association analysis for all cause dementia and Hypertension

	Gene	FHAT (p-value)	FHAT-O (p-value)	#variants	cumMAC
All cause dementia (N=129,670)					
	TREM2	5.2×10^{-6}	4.1×10^{-9}	45	4559
	PVR	1.2×10^{-5}	1.8×10^{-5}	75	2068
	EFCAB3	4.0×10^{-5}	4.2×10^{-5}	60	2579
	EMSY	4.4×10^{-5}	2.7×10^{-5}	158	1543
	KLC3	4.8×10^{-4}	1.4×10^{-5}	177	4174
	ABCA7	2.9×10^{-3}	4.1×10^{-5}	487	12179
Hypertension (N=129,206)					
	GATA5	4.1×10^{-5}	9.1×10^{-5}	88	5402
	FGD5	2.3×10^{-4}	4.3×10^{-5}	254	5269
	DDN	0.016	4.2×10^{-5}	107	1621

The exome-wide significance threshold is $\frac{0.05}{18,000} = 2.8 \times 10^{-6}$. The suggestive exome-wide significance threshold is $\frac{1}{18,000} = 5.6 \times 10^{-5}$. cumMAC is the cumulative minor allele frequency in the region. #variants is the total number of variants in the gene.

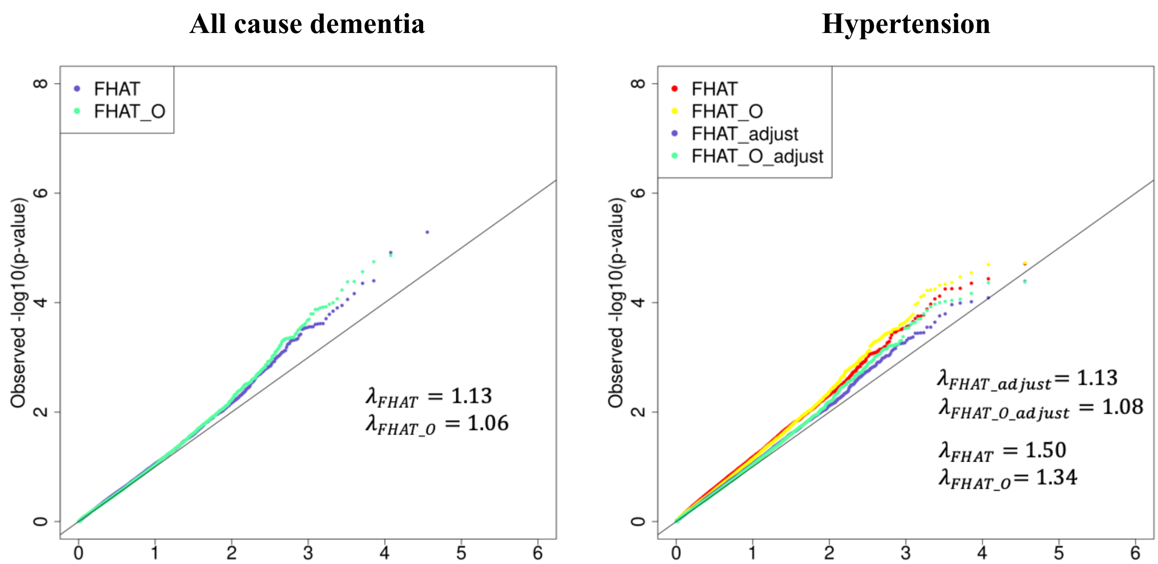

Figure 2.3 Quantile-Quantile plots of whole exome-wide analysis results for all cause dementia and hypertension

The p-values for regions with cumulative minor allele counts > 20 were used to generate the Quantile-Quantile plos. The left panel is the whole exome-wide analysis results for all dementia, where FHAT and FHAT-O were calculated using the model with the same covariates (age, sex, PC1-5, PC11) adjusted in AD/dementia known gene analysis. The right panel is the whole exome-wide analysis results for hypertension, where FHAT and FHAT-O were calculated using the model with the same covariates (age, age2, sex, BMI, PC1-PC5, PC8, and PC14) adjusted in hypertension known gene analysis. FHAT_adjust and FHAT-O_adjust were calculated from the adjusted hypertension analysis, where the spouse's hypertension status combining with other previously mentioned covariates were adjusted in the parental analysis.

2.5 Discussion

We proposed two novel approaches, FHAT and FHAT-O, that incorporate FH to increase power to detect rare variant associations in aggregation unit-based analysis. We also offered a novel way to adapt the LT-FH method to analyze rare variants. Because FH of disease is often collected through questionnaires in large cohorts, the added power is at no added cost. We applied our methods to exploit the FH from parents in simulation analysis and using the UK Biobank data, by assuming that the parents are conditionally independent.

We analyzed both parents through a single relatives' model, and combined the scores calculated from parents and probands with appropriate weights to calculate the test statistics. Because the probands' analysis is separate from the relatives' analysis, our methods can handle the missingness in FH as presented in (1) and (4), and one can include all probands with or without FH to optimize the usage of data.

The power was evaluated at alpha= 2.5×10^{-6} to represent the exome-wide significance for testing 20,000 genes as well as at a suggestive threshold of alpha= 2.5×10^{-5}. By assuming that the causal variants in older people have bigger effects compared to younger people, we showed FHAT and FHAT-O have greater power than SKAT-LTFH, SKATO-LTFH, with greatly reduced computational cost. We also note that, as we saw a slightly higher type I error inflation in SKAT-LTFH and SKATO-LTFH than FHAT and FHAT-O, we would expect more power gain in FHAT and FHAT-O when using an empirical significance level. Compared with SKAT and ACAT-V, FHAT has greater gain in power in most cases. However, FHAT and SKAT are less powerful than Burden tests and SKAT-O when there is a high proportion of causal variants, especially when the causal variants all have the same direction of effects. FHAT-O combines the features of both FHAT and FHAT-Burden, has robust power in many scenarios, and outperforms other methods, as shown in our extensive power simulations. ACAT-V has slightly higher power in some cases where the proportion of causal variants is low, which was expected because only a few genetic variants contribute to the results in ACAT-V, though the score statistic for FHAT and FHAT-O is calculated using a linear combination of squared scores from both causal and non-causal variants.

We further demonstrated that our methods have improved significance after incorporating FH from association analyses with all cause dementia and hypertension using genotypes and phenotypes collected from the UK Biobank. We compared results using FHAT, FHAT-O, SKAT-LTFH, and SKATO-LTFH for probands with both genotypes and phenotypes, and their parental history of disease to other methods only using probands. Variants in 8 known AD/dementia regions and 8 known hypertension regions were selected for the analysis. Using the significance level = 6.3×10^{-3} for testing 8 known genes, *BCL3* and *TOMM40* were significantly associated with all cause dementia while other known AD/dementia regions had improved significance compared to the methods that do not incorporate FH. Some of the hypertension genes were less significant using our method to incorporate FH, which might be caused by additional noise resulting from a similar hypertension prevalence in probands and their parents. The FHAT and FHAT-O approaches yielded similar conclusions compared to SKAT-LTFH, and SKATO-LTFH, respectively.

We evaluated type I error at various alpha levels and disease prevalence. We did not evaluate the type I error for SKAT-LTFH and SKATO-LTFH at the exome-wide significance (alpha= 2.5×10^{-6}) to limit the computational cost. The type I error of SKAT was previously found to be conservative when the disease prevalence is ~ 50%, and the Burden test was found to have appropriate type I error when the case-control ratio is balanced. [38,42,47] However, SKAT, SKAT-O, Burden and ACAT-V suffer from substantial inflated type I error when the prevalence is low, especially for lower alpha level (i.e., alpha < 2.5×10^{-4}). In contrast, the FHAT, SKAT-LTFH, FHAT-O and SKATO-LTFH control the type I error rates relatively better. The type I error is overall well controlled using FHAT and FHAT-O in most scenarios, but a high inflation occurs for alpha= $2.5\text{x}10^{-6}$ and prevalence = 10% where the number of cases and controls is unbalanced. Unbalanced case-control ratio yields inflated type I error rates because the imbalance invalidates the asymptotic assumption of logistic regression. Saddle point approximation (SPA) [10,13,31] method and efficient resampling (ER) [35] have been successfully used to calibrate binary phenotype based logistic mixed models [36] when case-control ratios are extremely unbalanced. In the future, we plan to adopt these cutting-edge methods to properly account for unbalanced case-control ratio.

In the exome-wide association analysis, we used the same covariates (age, sex, PC1-5, PC11) as we did in the known region analysis for all cause dementia. However, as the inflation was observed in our hypertension analysis (**Figure 2.3**), we further adjusted for the spouse's disease status in the parents' model to account for the correlations among parents due to household effects, in addition to the covariates of age, age^2, sex, BMI, PC1-PC5, PC8, and PC14. In the future, we will extend the current approaches to allow for correlation, as might be induced by household effect, in the analysis. Through the whole exome-wide analysis using FHAT and FHAT-O, we confirmed previously reported genes (*TREM2, PVR,* and *ABCA7*) [23,25,27] for AD/dementia as well as genes (*GATA5, FGD5,* *DDN*) [15,22,46] related to blood pressure and hypertension. Moreover, our methods identified three novel regions associated with all cause dementia (*EFCAB3, EMSY, KLC3*) using a suggestive exome-wide significance threshold. Replication analyses are needed to confirm these findings. While we observed inflated type I error for low prevalence in our simulations, we did not see evidence of large inflation of FHAT and FHAT-O in all cause dementia analysis, as seen from e Quantile-Quantile plot (**Figure 2.3**) and genomic control inflation factor (with $\lambda_{FHAT} = 1.13$ and $\lambda_{FHAT\_O} = 1.06$ for all cause dementia analysis). Although the method development, simulation studies and UK Biobank analysis described in the chaper were focusing on the population samples, our methods can also handle the ascertainment that happens in case-control analysis, because the likelihood can be written as the product of the retrospective proband information, taking ascertainment into consideration, and the (unascertained) relative likelihood: $P(G_i^P, Y_i^R | Y_i^P, X_i^P, X_i^R) = P(Y_i^R | G_i^P, Y_i^P, X_i^R) P(G_i^P | Y_i^P, X_i^P)$. The Equation (1) was derived based on the assumption of independence of the relatives' phenotype and probands' covariates conditional on the relatives' covariate and the strength of the associations in relatives. However, when the proband covariates are believed to have an effect on the relatives' disease status, one can adjust for such covariates in the relatives' model (3) to account for such an effect.

We demonstrated that FHAT and FHAT-O are computationally efficient methods compared to SKAT-LTFH and SKATO-LTFH. The significant reduced computational cost using FHAT and FHAT-O was showed in the analysis time to run 1000 aggregation unit-based tests. Although we focused on binary traits and rare variants, our method can be applied to analyze continuous traits using linear models and common variants. The framework in FHAT is flexible for various setting. While we applied FHAT and FHAT-O for probands with parental disease status available in simulations and the UK Biobank analysis, FHAT can be easily applied to other relative types. We also proposed an extension to FHAT, FHAT-O, to capture the features in SKAT-O, in particular the robustness of the power when all genetic variants have the positive effects and the proportion of causal variants is high. The framework can easily be extended to incorporate any other established aggregation unit-based based methods. Our methods that allow the incorporation of available FH are innovative compared to traditional rare variant studies that only use cases and controls, which have great potentials to promote genetic association research.

Chapter 3 Family History Aggregation Unit-based Tests in Family Studies with Application to the Framingham Heart Study

3.1 Introduction

Family-based study designs have been widely used in genetic association analysis, because these designs take advantage of similar environmental exposure, and additional quality control (QC) checks that can be performed in relatives. However, the standard genome-wide association studies (GWAS) suffer from inflated type I error if family correlation is not appropriately handled. While the GWAS suffers low power to detect rare variant associations with limited sample sizes, the aggregation unit-based test accessing the joint effect of variants in a region can increase power for rare variant analysis. Traditional aggregation unit-based tests such as Burden tests, [38] sequence kernel association test (SKAT) [59] , SKAT-O [37] and C-alpha [48] are only valid for studies with unrelated samples. Extensions to these existing methods for related individuals include burden test accounting for familial correlation (famBT) [6] family-based SKAT (famSKAT) [6] and MONSTER (Minimum P-value Optimized Nuisance parameter Score Test Extended to Relatives) [28] where the type I error is well controlled in family studies. We refer to MONSTER as famSKAT-O in this chapter.

Family history (FH) provides an overview of phenotypes among family members and is valuable for genetic association analysis. I developed a family history aggregation unit-based test (FHAT) to incorporate available FH into rare variant association analysis and demonstrated a power gain when exploiting additional data from relatives (Chapter 2). [57] FHAT was proposed to incorporate FH from a single relative or unrelated parents, and FHAT does not appropriately account for correlation among multiple related family members or related probands.

In this chapter, we propose a family-based FHAT (famFHAT) to adjust for familial correlation as a random effect through generalized linear mixed model (GLMM). The adjustment for relatedness allows famFHAT to preserve the correct type I error in studies containing related samples and prevents the lost in sample size that occurs when restricting analyses to unrelated samples in order to avoid false-positive results. We also develop an extended version of famFHAT, optimal FHAT (famFHAT-O). famFHAT-O combines the features of famFHAT and famFHAT-burden, which outperforms other methods in scenarios regardless of directions of genetic effects or proportions of causal variants. These methods can be applied to the instances where the data contain family samples either in probands with both genotypes and phenotypes or relatives with only phenotypes. For example, the approach is applicable when there are multiple siblings with both genetic and phenotype data and relatives with only phenotypes. The proposed methods, famFHAT and famFHAT-O, account for relatedness and can incorporate FH from relatives with different degrees of relationship with the probands. We demonstrate the validity of these approaches to incorporate multiple relatives from complex family structure through a simulation analysis, where the type I error is correctly controlled in famFHAT and famFHAT-O. After incorporating available FH from relatives, famFHAT and famFHAT-O are more powerful than famSKAT and famSKAT-O. The gene-based analyses of exome chip data from the Framingham Heart Study (FHS) yield improved significance in rare variant regions known to be associated with Alzheimer's disease (AD) and dementia using famFHAT and famFHAT-O compared to other standard methods that ignore FH. Furthermore, with enhanced power to detect rare variant associations, we identify novel genes AD, dementia, and T2D in the FHS data.

3.2 Methods

3.2.1 Accounting for Familial Correlation in Aggregation unit-based Tests

When related samples are present in a study, one can adjust for the correlation among observations through the GLMM with a random effect to obtain the family-based FHAT (famFHAT). Specifically, the GLMM accounting for correlation in probands can be specified as

$$g\left(E(Y_i^P \middle| G_i^P, X_i^P, \delta_i)\right) = X_i^P \alpha_P + G_i^P\, \beta_P + \delta_i,$$

where $g(\cdot)$ is the link function that connects the phenotype mean Y_i^P with the covariates vector X_i^P, the genotype vector G_i^P and the random effect δ_i for i proband. We assume the $n \times 1$ random effect vector δ that contains each entry δ_i follows the distribution of $N\left(0, \sigma_{G_P}^2 \Phi_\mathrm{P}\right)$, where Φ_P is the $n \times n$ matrix containing twice the kinship estimates obtained from family information or estimated from genotype data when available, and $\sigma_{G_P}^2$ is the parameter of variance component. [6] In this model, α_P is a vector of regression coefficients for covariate effects, β_P is a vector of regression coefficients for the observed genotypes in probands. The restricted maximum quasi-likelihood (REML) function for the

GLMM model is

$l_{REML} = -\frac{1}{2}\log|\Sigma_{\mathrm{P}}| - \frac{1}{2}\log\left|X^{P^T}\Sigma_{\mathrm{P}}^{-1}X^P\right| - \frac{1}{2}\left(Y^{P^*} - X^P\alpha_P\right)^T\Sigma_{\mathrm{P}}^{-1}(Y^{P*} - X^P\alpha_P)$, where X^P is the $n \times p$ covariate matrix, $Y^{P*} = (Y_1^{P*}, Y_2^{P*}, \dots, Y_n^{P*})$ is the working vector defined based on the type of phenotypes, and the Σ_{P} is the variance-covariance matrix (refer to Appendix B.1 for details). Under the null hypothesis,

$$\hat{\Sigma}_P = \hat{\sigma}_{G_P}^2\Phi_{\mathrm{P}} + \hat{R}_P^{-1},$$

where $\hat{R}_P^{-1} = \hat{\phi}_P^2\boldsymbol{I}$ for continuous traits and $\hat{R}_P^{-1} = diag\{1/(\hat{\mu}_{Pi}(1 - \hat{\mu}_{Pi}))\}$ for binary traits.

The genetic effect β_{P_j} for variant j is assumed to have an arbitrary distribution with mean zero and a variance of $w_j^2\tau$, where w_j is a pre-specified weight for variant j and τ is a variance component. Because testing $\beta = 0$ is equivalent to test $\tau = 0$, we take the derivative of REML respect to τ and fix other parameters estimated at the null hypothesis,

$$\frac{\partial l_{REML}}{\partial \tau}|_{\tau=0} = -\frac{1}{2}tr\left\{(\hat{\Sigma}_P^{-1} - \hat{\Sigma}_P^{-1}X^P\left(X^{P^{\mathrm{T}}}\hat{\Sigma}_P^{-1}X\right)^{-1}X^{P^{\mathrm{T}}}\hat{\Sigma}_P^{-1}\right)G^PWWG^{P^T}\}$$
$$+\frac{1}{2}\left(Y^{P^*} - X^P\hat{\alpha}_P\right)^T\hat{\Sigma}_P^{-1}GWWG^T\hat{\Sigma}_P^{-1}(Y^{P^*} - X^P\hat{\alpha}_P)$$
$$= -\frac{1}{2}tr\left\{PG^PWWG^{P^T}\right\} + \frac{1}{2}\left(Y^{P^*} - X^P\hat{\alpha}_P\right)^T G^PWWG^{P^T}\hat{\Sigma}_P^{-1}\left(Y^{P^*} - X^P\hat{\alpha}_P\right),$$

where $P = \hat{\Sigma}_P^{-1} - \hat{\Sigma}_P^{-1}X^P\left(X^{P^T}\hat{\Sigma}_P^{-1}X^P\right)^{-1}X^{P^T}\hat{\Sigma}_P^{-1}$ is the projection matrix. Because the first term is fixed and independent of the phenotype vector, the score statistic based on GLMM can be written as

$$Q_{GLMM} = \left(Y^* - X\hat{\alpha}\ \right)^T \hat{\Sigma}^{-1} GWWG^T \hat{\Sigma}^{-1} (Y^* - X\hat{\alpha}\ ),$$

which is equivalent to

$$Q_{famSKAT} = \frac{(Y^P - \hat{\mu}_P\ )^T G^P WWG^{P^T} (Y^P - \hat{\mu}_P\ )}{\hat{\phi}_P^2},$$

where $W = diag(w_1, w_2, \dots, w_m)$ is a pre-specified weight matrix with a size of m x m for the genotypes of m variants; G^P is the $n \times m$ genotype matrix with (i, v) element corresponding to the additively coded genotype for variant v of proband i ; $\hat{\mu}_P = (\hat{\mu}_{P_1}, \hat{\mu}_{P_2}, \dots, \hat{\mu}_{P_n})$ is the estimated mean of $Y^P = (Y_1^P, Y_2^P, \dots, Y_n^P)$ under H_0 of no genetic effect, $\hat{\phi}_P$ is the dispersion parameter estimate under H_0 and $\hat{\phi}_P = 1$ is fixed for binary traits.

Same to the Burden test, the famBT uses the same weights W to aggregate the rare variants, i.e., $\sum_{j=1}^{m} w_j g_{ij}$, and then test the association between this weighted sum of genotypes with the phenotype:

$$Q_{famBT} = \left[\frac{1}{\hat{\phi}_P} \sum_{i=1}^{n} (Y_i^P - \hat{\mu}_P\ ) \left(\sum_{j=1}^{m} w_j g_{ij}\right)\right]^2.$$

We combine the correlation-adjusted scores to formulate famFHAT and famFHAT-O using a weighted meta-analysis.

When we have FH from multiple relatives of probands, let $Y_j^{P'}$ and $G_j^{P'}$ denote the phenotype mean and genotype mean among the closest probands that are defined based on the known kinship coefficient for relative j, and X_j^R be the covariates in relatives. We evaluate the total association evidence through two separate GLMMs for probands and relatives:

$$g\left(E\left(Y_i^P \middle| G_i^P\ ,X_i^P\ ,\delta_i^P\right)\right) = X_i^P \alpha_P + G_i^P\ \beta_P + \delta_i^P, \qquad (1)$$

$$g\left(E\left(Y_j^R \middle| G_j^{P'}\ ,Y_j^{P'}, X_j^R\ ,\delta_i^R\right)\right) = X_j^R\ \alpha_R + G_i^{P'}\ \beta_R + Y_j^{P'} \lambda_R + \delta_i^R. \quad (2)$$

In the second equation, λ_R, a scalar, is the regression coefficient for probands' phenotypes for the relatives' model; α_R is a vector of regression coefficients for relatives' covariates; β_R is vector of regression coefficients for the m observed genetic variants in probands. Following the same steps to derive FHAT [57], the formula to calculate the score statistics of famFHAT to test the null hypothesis of no genetic effect is:

$$Q_{famFHAT} = \left[\frac{(Y^P-\hat{\mu}_P)^T}{\hat{\phi}_P} + \sum_k \frac{2\Omega_k D(R_k)(Y^{R_k}-\hat{\mu}_{R_k})^T}{\hat{\phi}_R}\right] G^P W\, W G^{P^T} \left[\frac{(Y^P-\hat{\mu}_P)}{\hat{\phi}_P} + \sum_k \frac{2\Omega_k D(R_k)(Y^{R_k}-\hat{\mu}_{R_k})}{\hat{\phi}_R}\right] \quad (3),$$

where $G^{P'}$ is $n \times m$ genotype matrix with the (j, v) element representing the genotype mean of variant v among the closest probands for relative j; Ω_k is the kinship coefficient for relative k. We assume that there is a superset of K relatives with the sample size of $n * K$, indexed k= 1 to K, that includes all possible types of relatives available on any of the probands. We define the n x n diagonal matrix $D(R_k)$ with ith diagonal element $d_i(R_k)$ to denote the availability of relative k's FH for proband i. Specifically, $d_i(R_k)$ equals to 0 if the values in Y^{R_k}, the phenotype vector for relative k of probands, are missing, and equals to 1 otherwise. Under the null hypothesis, $\hat{P}_P = \hat{\Sigma}^{-1}{}_P - \hat{\Sigma}^{-1}{}_P X_P \left(X_P^T \hat{\Sigma}^{-1}{}_P X_P\right)^{-1} X_P^T \hat{\Sigma}^{-1}{}_P$ and $\hat{P}_{R_k} = \hat{\Sigma}^{-1}{}_{R_k} - \hat{\Sigma}^{-1}{}_{R_k} X_{R_k} \left(X_{R_k}^T \hat{\Sigma}^{-1}{}_{R_k} X_{R_k}\right)^{-1} X_{R_k}^T \hat{\Sigma}^{-1}{}_{R_k}$ are the two projection matrices for proband and relative k, and $Q_{famFHAT}$ asymptotically follows a weighted sum of chi-squre distribution with df= 1, $\sum_{j=1}^{m} \lambda_j \chi_{1,j}^2$, where $\lambda_j s$ are the eigenvalues of $W\, G^{P^T}\left(\hat{P} + \sum_k 4\Omega_k^2 D(R_k) \hat{P}_{R_k} D(R_k)\right) G^P W$.

A unified test that combines famFHAT and famFHAT-Burden is

$Q_\rho = (1-\rho) Q_{famFHAT} + \rho Q_{famFHAT-Burden}$,

where

$$Q_{famFHAT-Burden} = \left[\frac{\sum_{i=1}^{n}\left(Y_i^P - \hat{\mu}_{Pi}\right)\left(\sum_{j=1}^{m} w_j g_{ij}\right)}{\hat{\Phi}_P} + \sum_k \frac{2\Omega_k \sum_{i=1}^{n} d_i(R_k)\left(Y_i^{R_k} - \hat{\mu}_{R_{k_i}}\right)\left(\sum_{j=1}^{m} w_j g_{ij}\right)}{\hat{\Phi}_R}\right]^2.$$

The value of ρ is selected to minimize the p-value, and we can write the famFHAT-O statistic as:

$$Q_{famFHAT-O} = \min_{0 \le \rho \le 1} P_\rho.$$

3.2.2 Incorporating FH from Multiple Relatives into Analyses

The GLMM can be used to account for correlation in the proband analysis or in the analysis of probands' relatives when we want to combine data from multiple relatives (e.g., relative k = 1 for mothers, k =2 for maternal grandfather, etc.) per proband. The famFHAT statistic in (1) is constructed by combining the scores from probands and relatives with appropriate weights. Because the score for un-genotyped relatives is down-weighted by twice of their kinship coefficient, the scores of relatives with different degrees of relatedness are combined by meta-analysis using relationship-appropriate weights, preventing adjustment for the correlation among relatives with different degrees of relatedness to the probands.

Therefore, we propose an alternative approach. Supposing that Ω is the scale vector with the size of $n * K$ containing the elements denoting kinship coefficient for each relative k, which is specified based on their relationship between probands, we can re-write the relatives' model (2) as

$$g\left(E\left(Y_i^R \middle| G_i^P\ Y^P, X_i^R\ ,\ \delta_i^R\right)\right) = X_i^R\ \alpha_R + 2\Omega\ *\ G_i^P\ \beta_P + Y_i^{P'} \lambda_R + \delta_i^R, (4)$$

based on fact that $\beta_P\ = 2\Omega\beta_R$. [55] Because model (2) is equivalent to (4), instead of performing related-specific analysis and appropriately down-weighting the score for each relative type, we can down-weight the genotypes based on their relationship with probands prior to computing the scores. Thus, the score statistics for probands and relatives can still be obtained through two separate analysis: probands analysis and a single analysis of probands' relatives. The scores from probands and their relatives are combined using meta-analysis to obtain famFHAT and famFHAT-O. Although it is an approximation, we have investigated this alternative approach through a simulation study. This allows the incorporation of FH from multiple relatives with different relatedness in the analysis through a single relatives' model. To address the situation where probands have multiple relatives, we take the following strategies to calculate our famFHAT statistic in the analysis:

1) For each relative, the genotype mean among the closest proband(s) are used in the score calculation, and the phenotype mean among the closest proband(s) are used as the covariate in the relatives' model; 2) We down-weight the genotypes based on the relationships with probands in the relatives' score calculation and then combine the down-weighted score with the probands' score to calculate famFHAT and famFHAT-O.

3.3 Simulation Studies

3.3.1 Validation of Incorporating FH from Multiple Relatives

3.3.1.1 Simulation Design

We first simulated genotypes for 5000 sibling pairs, and for one of the siblings in the pair, we simulated one offspring. The offspring in those 5000 families were treated as probands who have both available phenotypes and genotypes, and the proband in each simulated family has a parent and an aunt/uncle. The model used for the phenotype simulation is

$$\begin{pmatrix} Y^{aunt/uncle} \\ Y^{parent} \\ Y^{proband} \end{pmatrix} = 0.015 \begin{pmatrix} age^{aunt/uncle} \\ age^{parent} \\ age^{proband} \end{pmatrix} + 0.25 \begin{pmatrix} sex^{aunt/uncle} \\ sex^{parent} \\ sex^{proband} \end{pmatrix} + \begin{pmatrix} G_{causal}^{aunt/uncle} \\ sex_{causal}^{parent} \\ sex_{causal}^{proband} \end{pmatrix} \gamma + \varepsilon,$$

where $\varepsilon \sim MVN(0, \Sigma)$ with $\Sigma = \delta_g^2 \begin{pmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix} + \delta_e^2 I_{3\times3}$, and $\delta_g^2 = \delta_e^2 = 0.5$. In model, $age^{aunt/uncle}$, age^{parent} and $age^{proband}$ are vectors of continuous variable randomly selected from ages in UK Biobank data with a range from 38 to 72 for probands and 60 to 105 in parents; $sex^{proband}$ is a vector of binary variable generated from a Bernoulli distribution with probability for female = 56% in probands; $sex^{aunt/uncle}$ and sex^{parent} are the vectors of binary variables generated from the Bernoulli distribution with probability of having female= male= 50%; G_{causal}^{s} is the matrix containing the causal genotypes for individual s; and the γ is used to define the genetic effect [6]:

$$\gamma_j = \sqrt{\frac{c}{2MAF_j\left(1 - MAF_j\right)}}, (5)$$

where MAF_j is the MAF of causal variant j. Let R^2 denote the proportion of variance explained variants, V denote the vector representing the direction of genetic effects, and D denote the linkage disequilibrium (LD) correlation matrix. We conducted two sets of simulation studies, one to evaluate type I error and a second set of simulations to evaluate power. In the power analysis,

$$c = \frac{R^2}{V^T D V},$$

we fixed $R^2 = 2\%$ when all genetic effects are in the same direction and $R^2 = 3\%$ when half of genetic effects are in the same directions and half of the variants in the opposite directions. We simulated the data by assuming the proportion of causal variants = 50%. We set c equal to zero for the type I error analysis. On the liability scale, we converted the simulated continuous phenotypes to a binary trait with prevalence = 20%. In all models, we adjusted for age and sex, and we tested the regions containing 40 variants.

In the first analysis, we included all 5000 probands, the parents from a subset of 2500 families, and the aunts/uncles from the remaining 2500 families to create an unrelated relative subset. We calculated the scores for probands and relatives by fitting three separate logistic models:

$$logitP(Y_i^P = 1|G_i^P, X_i^P) = X_i^P \alpha_P + G_i^P \beta_P,$$

$$logitP\left(Y_i^{parent}\middle|G_i^P, X_i^{parent}\right) = X_i^{parent}\alpha_{parent} + Y_i^P \lambda_{parent} + G_i^P \beta_{parent},$$

$$logitP\left(Y_i^{aunt/uncle}\middle|G_i^P, X_i^{aunt/uncle}\right) = X_i^{aunt/uncle}\alpha_{aunt/uncle} + Y_i^P \lambda_{aunt/uncle} + G_i^P \beta_{aunt/uncle}.$$

Because we have unrelated relatives in this case, the FHAT method [57] using the standard logistic model is valid to apply to calculate the scores. We down-weight the relative scores by twice of kinship coefficient (i.e., ½ of the score for parents and ¼ of the score for aunts/uncles), and combine them with the score from probands using a weighted meta-analysis framework to calculate FHAT1, i.e.,

$$Q_{FHAT_1} = \left[\frac{(Y^P - \hat{\mu}_P)^T}{\hat{\phi}_P^2} + \frac{(Y^{parents} - \hat{\mu}_{parents})^T}{2\hat{\phi}_{parents}^2} + \frac{(Y^{aunts/uncles} - \hat{\mu}_{aunts/uncles})^T}{4\hat{\phi}_{aunts/uncles}^2}\right]$$

$$G^P W\, W G^{P^T}\left[\frac{(Y^P - \hat{\mu}_P)}{\hat{\phi}_P^2} + \frac{(Y^{parents} - \hat{\mu}_{parents})}{2\hat{\phi}_{parents}^2} + \frac{(Y^{aunts/uncles} - \hat{\mu}_{aunts/uncles})}{4\hat{\phi}_{aunts/uncles}^2}\right],$$

where G^P is the observed genotype matrix in 5000 probands; $\hat{\mu}_P$, $\hat{\mu}_{parents}$ and $\hat{\mu}_{aunts/uncles}$ are the estimated means of Y^P, $Y^{parents}$ and $Y^{aunts/uncles}$ under H_0, dispersion parameters are fixed to 1 for binary traits.

In the second analysis, we combined 2500 parents and 2500 aunts/uncles into a single relatives' model. Specifically, the standard logistic model for probands, and logistic model for relatives (parents and aunts/uncles) were used,

$$logitP\left(Y_i^P = 1\middle|G_i^P, X_i^P\right) = X_i^P \alpha_P + G_i^P \beta_P,$$

$$logitP\left(Y_i^{relative}\middle|G_i^P, X_i^{relative}\right) = X_i^{relative}\alpha_{relative} + Y_i^P \lambda_{relative} + G_i^P \beta_{relative} + \delta_i^{relative}.$$

We down-weighted the genotypes for parents and aunts/uncles by 1/2 and 1/4, respectively, and we combined the corresponding scores for probands and relatives (parents and aunts/uncles) through meta-analysis to calculate FHAT_2,

$$Q_{FHAT_2} = \left[\frac{(Y^P - \hat{\mu}_P)^T}{\hat{\phi}_P^2} G^P W + \frac{(Y^{relatives} - \hat{\mu}_{relatives})^T}{\hat{\phi}_{relatives}^2} 2\Omega\, G^P W\right]$$

$$\left[W G^{P^T} \frac{(Y^P - \hat{\mu}_P)}{\hat{\phi}_P^2} + 2\Omega\, W G^{P^T} \frac{(Y^{relatives} - \hat{\mu}_{relatives})}{\hat{\phi}_{relatives}^2}\right],$$

where $\hat{\mu}_P$ and $\hat{\mu}_{relatives}$ are the estimated means of Y^P and $Y^{relatives}$ under the null hypothesis, respectively.

3.3.1.2 Simulation Results

We assessed the type I error rates by performing 10,000 simulation replicates, see **Table 3.1**. The power for different scenarios of genetic effect directions using 1000 replicates was evaluated (**Table 3.2**). **Figure 3.1** summarizes the p-values computed by FHAT_1 and FHAT_2 from two set of analysis described in previous section. Because the results were comparable between the approach analyzing different relative types separately versus the approach combining all relatives in the analysis by down-weighting genotypes proportionally to the coefficient of relationship with the proband, we concluded that it is feasible to analyze all relatives with various degrees of relationship through a single regression model and down-weight the probands' genotypes in their score calculation.

Table 3.1 Type I error rates for $FHAT_1$ and $FHAT_2$

Alpha	Type I Error for $FHAT_1$	Type I Error for $FHAT_2$
0.05	0.052	0.053
0.01	0.0094	0.0094
0.001	0.001	0.0011
0.005	0.0048	0.0050

Table 3.2 Power for $FHAT_1$ and $FHAT_2$

Genetic Effects (+/-)	Alpha	Power for Analysis 1	Power for Analysis 2
(100/0)	10^{-4}	0.956	0.953
	10^{-5}	0.89	0.885
	10^{-6}	0.801	0.789
	10^{-7}	0.698	0.695
	10^{-8}	0.581	0.565
(50/50)	10^{-4}	0.716	0.716
	10^{-5}	0.537	0.532
	10^{-6}	0.391	0.388
	10^{-7}	0.277	0.279
	10^{-8}	0.217	0.214
+/-/ indicates the proportion of variants with positive, negative effects			

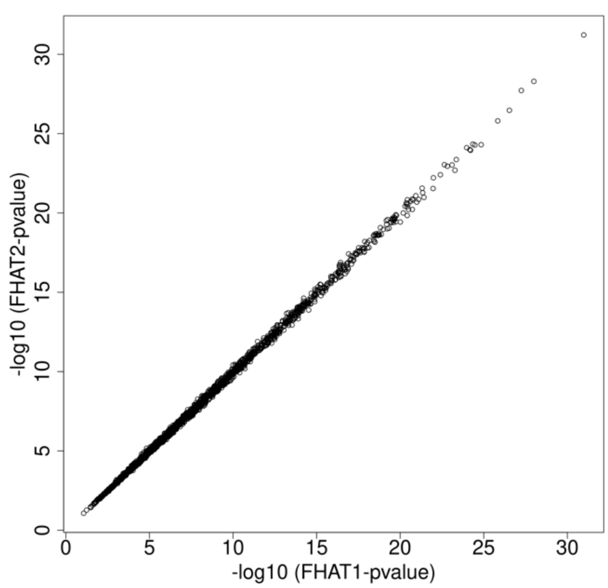

Figure 3.1 Comparison of p-values calculated using $FHAT_1$ and $FHAT_2$

3.3.2 Complex Family Structure

3.3.2.1 Simulation Design

To evaluate the performance of famFHAT and famFHAT-O combining FH from multiple relatives with various relationships with probands, we simulated 400 families containing 18 family members from 3 generations (**Figure 3.2**) to have probands and relatives with complex family structure.

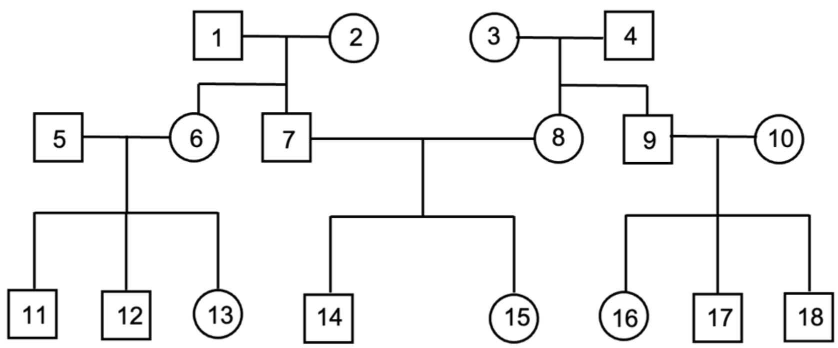

Figure 3.2 Pedigree used for complex family structure simulation

We randomly assigned two haplotypes simulated from HapGen2 [54] to each of the founders. The offspring were randomly assigned one haplotype without recombination from each parent. The model that we used for simulating the continuous phenotypes is specified as follows,

$$\begin{pmatrix} Y^{s=1} \\ Y^{s=2} \\ \dots \\ Y^{s=18} \end{pmatrix} = 0.015 \begin{pmatrix} age^{s=1} \\ age^{s=2} \\ \dots \\ age^{s=18} \end{pmatrix} + 0.25 \begin{pmatrix} sex^{s=1} \\ sex^{s=2} \\ \dots \\ sex^{s=18} \end{pmatrix} + \begin{pmatrix} G_{causal}^{s=1} \\ G_{causal}^{s=2} \\ \dots \\ G_{causal}^{s=18} \end{pmatrix} \gamma + \varepsilon,$$

where $\varepsilon \sim MVN(0, \Sigma)$; $\Sigma = 0.4 * \Phi + 0.6 * I_{18\times18}$, is the covariance matrix and Φ is twice the kinship matrix representing the relationship among the 18 family members; Y^s, age^s, sex^s are the values for phenotype, age, sex of individual s in the pedigree (**Figure 3.2**);

G_{causal}^{s} is the matrix containing the causal genotypes for individual s; and the γ that controls the genetic effects is defined in (5). We fixed R^2 to 2% in the power analysis when all causal variants have deleterious effects, and we increased this value to 3% for the scenario where half of the causal variants have deleterious effects and half of the causal variants have deleterious effects on the liability scale for binary trait. The continuous phenotypes were converted to binary traits using the normal approximation to obtain a pre-specified prevalence on the liability scale for binary trait.

In the simulations, we randomly selected 9 probands from each family samples contributing both genotypes and phenotypes and the remaining 9 samples from each family contribute only phenotypes. We repeated this process randomly for each family to have random and complex relationships among probands and relatives. Therefore, we had 3,600 probands and 3,600 relatives included in the analysis. Only the relatives who have at least one proband were used in the analysis. The regions with 30 variants were analyzed.

In the analysis of famFHAT and famFHAT-O, we used both simulated phenotypes and genotypes for probands, whereases only phenotypes were used in the form of FH for relatives by assuming their genotypes were missing. We compared the results of type I error rates and power to famSKAT and famSKAT-O that only used proband data.

3.3.2.2 Simulation Results

A total of 10,0000 replicates were generated to evaluate the type I error at different alpha levels. The results **in Table 3.3** show the type I error rates of famFHAT and famFHAT-O using both probands' disease status and FH from relatives, and famSKAT and famSKAT-O using only probands' data for prevalence = 20% and 30%. We also compared the family-based methods to the methods that do not account for correlation in family samples. We demonstrated that famFHAT, famFHAT-O, famSKAT, and famSKAT-O greatly improve the control of the type I error rates compared to their unadjusted versions that ignore the family structure. We noted that famFHAT and famFHAT-O have slightly higher inflation compared to famSKAT and famSKAT-O, which might be caused by the residual correlation among scores in probands and relatives.

Table 3.3 Type I error rates of famFHAT, famSKAT, famFHAT-O, famSKAT-O, FHAT, SKAT, FHAT-O and SKAT-O in family study

Alpha	**famFHAT**	**famSKAT**	**famFHAT-O**	**famSKAT-O**	**FHAT**	**FHAT-O**	**SKAT**	**SKAT-O**
Prevalence = 30%								
0.1	1.1	1.0	1.1	1.0	1.6	1.5	1.5	1.4
0.05	1.1	1.0	1.1	1.0	1.8	1.7	1.6	1.5
0.01	1.1	0.9	1.3	1.0	2.3	2.2	1.7	1.8
0.005	1.2	0.9	1.4	1.1	2.3	2.5	1.8	2.0
5×10^{-4}	1.2	1.0	1.9	1.6	3.2	4.1	2.7	3.5
Prevalence = 20%								
0.1	1.1	1.0	1.0	0.9	1.6	1.4	1.4	1.3
0.05	1.1	1.0	1.1	1.0	1.7	1.6	1.5	1.4
0.01	1.2	1.0	1.3	1.1	2.0	2.1	1.8	1.9
0.005	1.2	1.0	1.5	1.2	2.3	2.5	2.0	2.2
5×10^{-4}	1.3	1.4	1.9	2.0	3.3	4.1	3.3	4.1

The number in each cell represents the ratio of type I error and expected significance level (column 'Alpha'). Type I error was evaluated from the proportion of p-values less than or equal to each corresponding alpha level. All tests used the same Wu weights with beta (MAF_j; 1, 25). The analyses were restricted to rare variants with MAF < 1%.

Since the unadjusted methods (FHAT, FHAT-O, SKAT, and SKAT-O) have incorrect type I error in analysis of family samples, we did not include them in the power simulations. We estimated the power for 3600 probands with available FH from 3600 relatives with different degrees of relationship for a prevalence= 20%. The power was evaluated for alpha= 5×10^{-4} as it was the minimum alpha we tested in the type I error simulation. As indicated in **Figure 3.3**, our methods (famFHAT and famFHAT-O) outperform famSKAT and famSKAT-O after incorporating family history from parents in all scenarios.

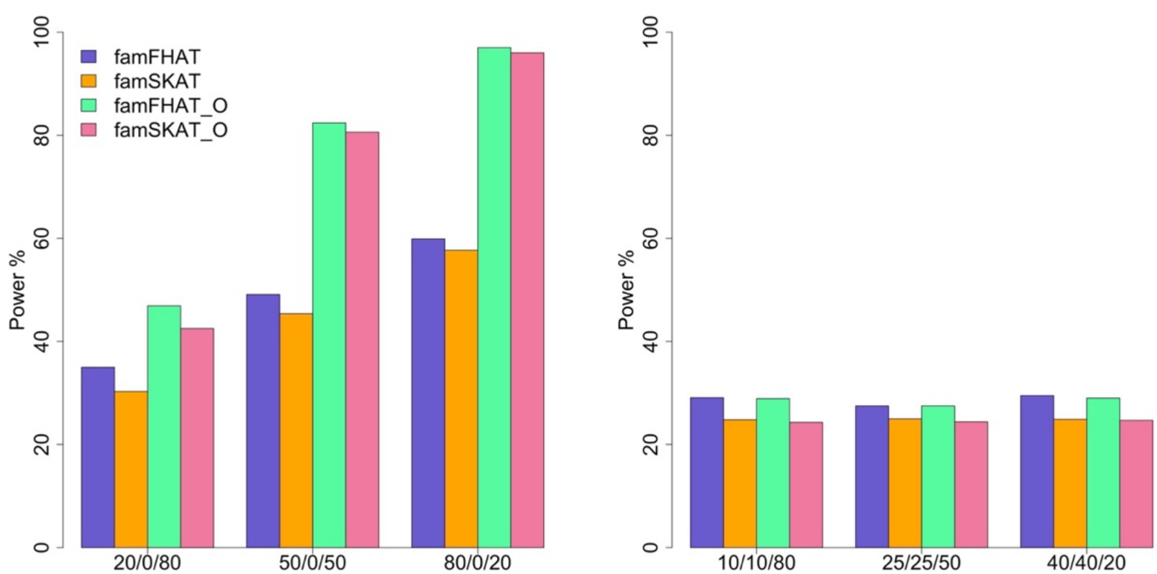

Figure 3.3 Empirical power of famFHAT, famFHAT-O, famSKAT, and famSKAT-O

In each plot, the x axis in the format of +/-/0 indicates the proportion of variants with positive, negative and no effects. Each bar shows the empirical power evaluated as the proportion of p-values less than or equal to alpha= 5×10^{-4}. The analyses were restricted to rare variants with MAF < 1%. The disease prevalence was set to 20%. All methods used the same Wu weights with beta (MAF_j; 1, 25). famFHAT, famFHAT-O analyzed 3,600 probands and incorporated FH from 3,600 relatives, while famSKAT and famSKAT-O only included data for probands.

3.4 Application to the Framingham Heart Study

3.4.1 Analysis of Exome Chip Data

The Framingham Heart Study (FHS) is a community-based, longitudinal cohort study. The cohort comprises residents of Framingham, Massachusetts, and these residents have undergone up to 32 examinations, performed every 2 years, that have involved detailed history taken by a physician, a physical examination, and laboratory testing. [12] The participants in the cohort's offspring study have completed up to 9 examinations, which have taken place approximately every 4 years. [29] The diagnosis of AD and related dementia in the FHS is based on criteria for possible, probable, or definite AD from the National Institute of Neurological and Communicative Disorders and Stroke and the AD and Related Disorders Association (NINCDS–ADRDA). [2, 45] The diagnosis of diabetes is based on a fasting plasma glucose > 125 mg/dl or non-fasting plasma glucose > 200 mg/dl or HbA1c Hemoglobin A1c (HbA1c) > 6.5 or history of diabetes treatment such as insulin or an oral hypoglycemic agent.

We applied our methods (famFHAT and famFHAT-O) to analyze exome chip data on GRCh37/hg19 in the FHS to investigate the regions associated with three phenotypes: AD, dementia, and T2D. In the FHS analysis, we assumed that probands are those who have both available genotypes and phenotypes, and relatives are those who with available phenotypes only (i.e., FH). The relatives were defined as the phenotyped but not genotyped samples who were related to at least one proband based on the kinship matrix. We used famFHAT and famFHAT-O to combine all FHS relatives without available genotypes but available phenotype data into an analysis to evaluate the association between variant sets or genes and disease status. Note that the inclusion of probands did not depend on the availability of FH from their relatives, since the probands' analysis was separate from the relatives' analysis. We classified coding variants for aggregation unit-based testing for exome chip analysis, which include nonsynonymous exonic variants, splicing exonic variants, splicing nonsynonymous exonic variants, exonic splicing stop gained, exonic splicing stop lost, splicing, stop gained and stop lost. We focused on investigating the associations of rare variants with MAF < 1%.

As shown in **Table 3.4**, a total of 3,949 probands with available both AD/dementia status and exome chip genotypes (P= 11.3% in AD, P= 14.3% in dementia), and 1,744 relatives with available dementia/AD status but missing genotypes were involved in the dementia and AD analyses (P= 14.3% in AD, P= 19.6% in dementia). We adjusted for age at the time of DNA draw, education status, and sex in AD analyses. Our dementia analysis used the same covariates as for AD analysis.

The T2D analysis included 7,356 probands with available T2D status and genotypes (P= 13.1%), as well as 2,765 ungenotyped relatives with available T2D status (P= 16.8%). We adjusted T2D analysis for age at the last exam, BMI at the last exam, and sex. We first tested 3 previously reported rare variant genes for AD and dementia: *APOE*, *SLC24A4*, and *INPP5D*, and 3 known genes for T2D: *PSD4, MRPL46 and GPD2,* and then explored novel regions using all exome chip data.

Table 3.4 Disease prevalence in the FHS

	Probands (with both genotypes and phenotypes)	**Relatives** (with phenotypes)
AD		
Sample size	3949	1744
Disease prevalence	11.3%	14.3%
Dementia		
Sample size	3949	1744
Disease prevalence	14.3%	19.6%
T2D		
Sample size	7356	2765
Disease prevalence	13.1%	16.8%

3.4.2 Results

The family structures in FHS are very complex, and probands could have one relative, both parents, and/or multiple relatives. The degrees of relatedness were calculated based on known family relationships using Kingship2 R package. [52] **Table 3.5.** summarizes the types of relationships between relatives and probands among all FHS families.

Table 3.5 Relationships between probands and relatives in the FHS

Relationship between proband and relatives	**Kinship Coefficient**	**Example Type**
First degree	0.25	parents / daughter / son
Second degree	0.125	grandparents/ grandchildren /aunt / uncle / niece / nephew
Third degree	0.0625	first cousin
Fourth degree	0.03125	great-great-grandparents/ great-great-grandchildren

We selected variants with MAF < 1% and MAC >1 available on the exome chip. When summarizing the results, we restricted the genes with cumulative MAC (cumMAC) > 20 in probands (**Table 3.6, Table 3.7, and Figure 3.4**). A total of 8,218 genes and 6,831 genes were selected to summarize the results of T2D and AD/dementia, respectively. The top genes in the table were selected using suggestive significance thresholds, which were calculated by 1 divided by the number of tested genes with cumMAC > 20 in probands (i.e., p-value < $1.5x10^{-4}$ for AD/dementia, p-value < $1.2x10^{-4}$ for T2D). No associations were detected between these three traits that using stricter multiple-testing corrected thresholds (p-value < $7.3x10^{-6}$ for AD/dementia, 0.05/6,831; p-value < $6.1x10^{-4}$ for T2D, 0.05/8218). The novel findings on the suggestive cut-offs will need to be validated from replication analysis or functional studies.

From investigating associations between previously identified genes for AD/dementia [3], we identified that *APOE* and *SLC24A4* had improved significance using famFHAT and famFHAT-O, compared to famSKAT and famSKAT-O, respectively, for both AD and dementia (**Table 3.7**). The gene *INPP5D* had a more significant p-value using famFHAT compared to famSKAT for AD, after incorporating additional data from disease status from relatives. From applying the methods to test previously reported rare variant regions for T2D based on previous findings, [58] two genes (*MRPL46* and GPD2) showed an improved significance using famFHAT compared to famSKAT. *MRPL46* had the smaller p-value estimated from famFHAT-O after incorporating T2D status from relatives compared to the analysis of probands alone using famSKAT-O. (**Table 3.8**).

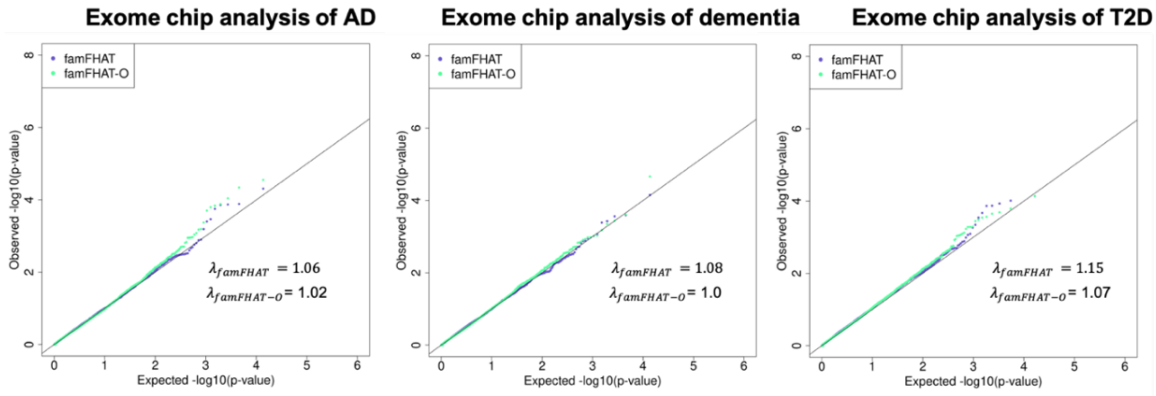

Figure 3.4 Quantile-Quantile plots from FHS exome chip analysis for AD, dementia and T2D

Table 3.6 Exome chip analysis of AD and dementia

Gene	#variants	cumMAC	famFHAT (p-value)	famFHATO (p-value)
AD				
CYP26B1	3	20	8.4×10^{-5}	8.8×10^{-5}
CCR5	7	70	4.4×10^{-4}	9.9×10^{-5}
ODZ3	11	92	9.7×10^{-5}	1.2×10^{-4}
KIAA0368	9	108	2.2×10^{-4}	3.6×10^{-5}
PYGM	11	133	8.4×10^{-5}	7.4×10^{-5}
ZSCAN18	2	38	7.6×10^{-5}	7.0×10^{-5}
Dementia				
KIAA0368	9	108	7.1×10^{-5}	2.2×10^{-5}

The suggestive significance threshold is $\frac{1}{6{,}831} = 1.5\text{x}10^{-4}$. cumMAC is the cumulative minor allele counts in probands for the gene we tested. #variants is the total number of variants tested in the gene.

Table 3.7 Exome chip analysis of T2D

Gene	#Variants	cumMAC	famFHAT	famFHATO
PDE8A	4	22	4.2×10^{-3}	1.1×10^{-4}
MAP3K3	3	29	8.7×10^{-4}	1.2×10^{-4}

The suggestive significance threshold is $\frac{1}{8{,}218} = 1.2\text{x}10^{-4}$. cumMAC is the cumulative minor allele counts in probands for the gene we tested. #variants is the total number of variants tested in the gene.

Table 3.8 Investigating association between traits and reported genes

Gene	#variants	cumMAC	famFHAT (p-value)	famSKAT (p-value)	famFHAT-O (p-value)	famSKAT-O (p-value)
T2D						
PSD4	13	100	0.96	0.88	0.77	0.82
MRPL46	4	94	0.84	0.70	0.86	0.83
GPD2	6	61	0.39	0.54	0.14	0.12
AD						
APOE	2	24	0.16	0.35	0.22	0.46
SLC24A4	10	109	0.72	0.80	0.23	0.35
INPP5D	10	95	0.72	0.83	0.44	0.11
Dementia						
APOE	2	24	0.09	0.28	0.12	0.37
SLC24A4	10	109	0.57	0.72	0.10	0.23
INPP5D	10	95	0.73	0.68	0.34	0.10

cumMAC is the cumulative minor allele counts in probands for the gene we tested. #variants is the total number of variants tested in the gene.

3.5 Discussion

In this work, we proposed two novel approaches, famFHAT and famFHAT-O, to account for correlation among family members when incorporating FH information in aggregation unit-based tests. We adapted the GLMM to analyze related probands with both genotypes and phenotypes or related relatives with phenotypes. The estimates that used for score calculations were obtained from the R package GMMAT. [8] We presented the methods using the GLMM. famFHAT and famFHAT-O are flexible methods and can be applied to both continuous traits and binary traits by using the appropriate link function.

We showed that we obtained correct type I error rates, and similar results compared to the scores that were calculated by combining the weighted-scores from relatives, when combining relatives with different degrees of relationship with the proband into a single model by down-weighting their genotypes before calculating the scores. This demonstration allows us to combine available FH from multiple relatives into aggregation-based association test so that the methods are not restricted to exploit FH from a single relative. We noted that one limitation is that this approach relies on the known relationship, and it cannot be implemented for unknown relationship among relatives using an empirical kinship matrix, because the empirical kinship matrix needs to be estimated using known genotypes from probands and relatives. When the proband has FH from multiple relatives, we took the genotype mean and phenotype mean among the closest proband in their score calculation. We would expect that the incorporation of FH from close relatives yields more power gain than distant relatives. Moreover, when relatives are related to multiple probands, there might be some residual correlation between the proband and the relative scores which could result in inflated type I error rates.

In the simulation studies (Section 2.3), we showed that famFHAT and famFHAT-O have appropriate type I error rates in the presence of correlation through a simulation study, and maintain greater power than famSKAT after combining additional phenotype data from relatives. However, the unadjusted method that ignore correlation among family samples suffer from the substantial inflated type I error. Because famFHAT-O combines the features of famFHAT and famFHAT-burden, it has the most robust power in all scenarios we've tested.

As an illustration, we applied our method to analyze AD, dementia, and T2D using FHS exome chip data. famFHAT and famFHAT-O incorporated all available disease status data from relatives. Most of the known rare variant regions for AD, all dementia, and T2D were shown to have improved significance using our methods compared to the methods without adjustment of FH. By testing all regions using the exome chip data, we found novel and suggestive regions for AD, dementia, and T2D using our methods. However, those novel findings have not been validated. While the type I error is well controlled using famFHAT and famFHAT-O in family studies, there is no need to limit the analysis to an unrelated set thus maintaining a large sample size in most studies. By further incorporating FH, these methods boost power to detect rare variant associations in studies with correlated data.

Chapter 4 Robust Family History Aggregation Unit-based Methods for Unbalanced Case-control Designs

4.1 Introduction

Many studies for binary phenotypes have extremely unbalanced case-control ratios, due to the low prevalence of diseases in cohort studies such as biobanks. The generalized linear mixed effect models (GLMMs) have been widely used to solve the issues caused by relatedness and population structure for both quantitative and binary traits in the genetic association analysis. However, GLMM suffers from inflated type I error in unbalanced study designs and the inflation affects mostly very rare variants because of the inaccurate asymptotic distribution used to evaluate the significance of the score statistic for binary traits. [1]

The saddle point approximation (SPA) [31] method and efficient resampling (ER) [35] have been successfully used to calibrate the distribution of score statistics for binary phenotype relying on GLMM when case-control ratios are extremely unbalanced. SAIGE and SAIGE-gene have been recently proposed to accurately adjust for the case-control imbalance for binary trait, where the p-values are calculated using SPA method through the cumulant-generating function (CGF).

Previously, we demonstrated that incorporating additional data in the form of family history (FH) from relatives in genetic association analyses can enhance the statistical ability to identify rare variant associations by dramatically increasing sample size. Additionally, methods that incorporate FH control type I error better in unbalanced studies compared to the standard methods that ignore the FH. [57] However, the inflation of type I error is still a concern for lower alpha level and extremely low disease prevalence. As an extension to our previously developed methods that enable the incorporation of FH in unrelated samples and related samples, in this chapter, we propose robust versions to provide accurate results for extremely low disease prevalence in either probands or relatives for binary traits, while maintaining optimal power to detect rare variant associations.

4.2 Methods

4.2.1 GLMM-based Individual Variant Score

For a study comprised of n probands with available genotypes for m variants, we let Y_i^P and X_i^P denote the disease status and covariate matrix for proband i, respectively, and a nxm matrix G_i^P with the element g_{ij} representing the additively coded genotypes for variant j of probands i. The GLMM for the binary trait can be specified through a logit link function:

$$logit\left(P\left(Y_i^P = 1 \middle| G_i^P, X_i^P, \delta_i^P\right)\right) = X_i^P \alpha_P + G_i^P \beta_P + \delta_i^P, (1)$$

where α_P is a coefficient vector for covariate effects, β_P is a coefficient vector for the observed genotypes, δ_i^P is the random effect that is assumed to follow a normal distribution $N(0, \sigma_{G_P}^2 \Phi_\mathrm{P})$ where Φ_P is the $n \times n$ matrix containing relatedness estimates calculated from pedigree or estimated form available genotypes and $\sigma^2_{G_P}$ is the genetic variance parameter. When the study does not contain related samples, we set $\delta_i^P = 0$ and use the standard logistic model instead. Let W be the $m \times m$ diagonal matrix with the diagonal element w_j representing the pre-specified weight for variant j, then the genetic effect β_P follows a distribution with mean $W\boldsymbol{I}_m\beta$ and covariance τW^2, where τ is the variance component parameter and $\boldsymbol{I}_m$ is the all-ones vector with length of m. The score statistic for variant j based on probands' model (1) can be determined (where the dispersion parameter $\phi_P = 1$):

$$S_{P_j} = \sum_{i=1}^{n} g_{ij}(Y_i^P - \hat{u}_i^P).$$

In the above score, $\hat{u}_i^P$ is the estimated probability of having the disease that takes account the correlation structure among individuals using the null model of no genetic effect:

$$\hat{u}_i^P = logit\left(P_0\left(Y_i^P = 1 \middle| X_i^P, \delta_i^P\right)\right) = X_i^P \alpha_P + \delta_i^P.$$

Under the null hypothesis, $S^P = \left(S_{P_1}, S_{P_2}, \dots, S_{P_m}\right)$ follows a multivariate normal distribution $MNV(0, G^P \hat{P} G^P)$, where $\hat{P} = \hat{\Sigma}^{-1}{}_P - \hat{\Sigma}^{-1}{}_P X_P \left(X_P^T \hat{\Sigma}^{-1}{}_P X_P\right)^{-1} X_P^T \hat{\Sigma}^{-1}{}_P$ with $\hat{\Sigma}_P{}^{-1} = \hat{\sigma}^2_{G_P} \Phi_P + diag\{1/(\hat{\mu}_{Pi}(1 - \hat{\mu}_{Pi}))\}$.

4.2.2 Aggregation unit-based Test Statistics with SPA and ER

The normal approximation performs well when the score statistic S_j is within two standard deviations of the mean. [13] However, the skewed distribution of the score statistic when case-control ratio is extremely unbalanced can result in inflation in type I error rates. SPA is an approach for p-value calculation using all cumulants through the cumulant generating function (CGF), which can provide a more accurate approximation of the score distribution compared to the normal approximation using only two cumulants. [13,31,10] However, due to the poor performance for low minor allele count (MAC), the ER is the preferred method that can provide more accurate p-value for very rare variants (i.e., MAC $\leq$ 10) by resampling the disease status among individuals with low MAC in any variants. The details of SPA and ER have been discussed elsewhere. [60, 62] Here we took a similar approach to the one described in Zhou et al. [62] Let $\chi^2_{quantile}$ be the quantile function for the chi-square distribution with one degrees of freedom (df= 1), the variance for $S^2_{P_j}$ can be adjusted using the following formula:

$$\tilde{V}_j^P = \frac{S^2_{P_j}}{\chi^2_{quantile}\left(1 - \ \tilde{p}_j\right)}, (2)$$

such that the p-value for variant j is the same as $\tilde{p}_j$, that is, the p-value estimated using SPA (when MAC > 10) or ER (when MAC <= 10). Given the fact that SPA performs better for the single variant tests, a further adjustment was made using Burden tests because they can be written as the square of the sum of single variant scores across all samples, i.e., $Q_{P_{Burden}} = \left(S_{P_{Burden}}\right)^2$ with $S_{P_{Burden}} = \sum_{i=1}^{n} \sum_{j=1}^{m} w_j g_{ij} \left(Y_i^P - \hat{u}_i^P\ \right)$. We define $\tilde{r}^P = \min\left(1, (w^T \tilde{V}^{P^{\frac{1}{2}}} D \tilde{V}^{P^{\frac{1}{2}}} w)/\tilde{V}^P_{Burden}\ \right)$, where $\tilde{V}^P = \left(\tilde{V}_1^P, \tilde{V}_2^P, \dots, \tilde{V}_m^P\right)$, and $\tilde{V}^P_{Burden} = \left(\tilde{V}^P_{Burden_1}, \tilde{V}^P_{Burden_2}, \dots, \tilde{V}^P_{Burden_m}\ \right)$is the adjusted variance for ${S^2}_{P_{Burden}}$ using (2), $w = (w_1, w_2, \dots, w_m)$ is the vector of weights for m variants, and D is the correlation matrix of m variants. We use the calibrated distribution when the case-control ratio is unbalanced to derive an aggregation unit-based test,

$$\tilde{S}_P \sim MVN\,(0{,}0{,}\left(\frac{\tilde{V}^P}{\tilde{r}^P}\right)^{\frac{1}{2}} D \left(\frac{\tilde{V}^P}{\tilde{r}^P}\right)^{\frac{1}{2}}). (3)$$

4.2.3 Robust Methods to exploit FH in Aggregation unit-based Test in Unbalanced Designs

We have demonstrated the important contribution of FH in genetic association studies in recently developed methods that exploit FH for rare variant associations. [57] The methods incorporate FH as a phenotype through a separate relatives' model into aggregation unit-based test and evaluate the total association through a weighted meta-analysis combining the evidence from probands and relatives. These methods successfully enhance power to detect disease loci especially for late-onset diseases where the disease has low prevalence in younger probands. The methods of FHAT and optimal FHAT (FHAT-O) were proposed to incorporate FH from unrelated relatives (a single relative or two unrelated parents), and will have inflated type I error in studies with family samples when ignoring familial correlation. In this regard, family-based methods (famFHAT and famFHAT-O) using the GLMM that correctly adjust for correlation among family members using a random effect were developed in Chapter 3.

For a binary trait, the probands' analysis can be performed based on model (1). When the case-control ratios is extremely unbalanced, we can use $\tilde{S}^P$ specified in (3) with adjusted variance to calculate the score in probands, and we can use the same strategy to adjust the variance for relatives who only contribute phenotype data. In the notations, we first assume that all probands have available FH from a superset of K relatives containing all possible types of relatives within a sample size of $n * K$. Based on the model to analyze FH from relatives,

$$logit\ \left(P\left(Y_i^R = 1 \middle| G_i^P,\ X_i^R,\ Y_j^P\ , \delta_i^R\right)\right) = X_i^R \alpha_R + G_i^P\ \beta_R + Y_j^P\ \lambda_R + \delta_i^R, (4)$$

the unadjusted score statistics are

$$S_{R_{k_j}} = \sum_{i=1}^{n} d_i(R_k) g'_{ij} (Y_i^{R_k} - \hat{\mu}_{R_{k_i}}),$$

$$S_{Burden}^R = \sum_{i=1}^{n} d_i(R_k) {g'}_i^{Burden} (Y_i^{R_k} - \hat{\mu}_{R_{k_i}}),$$

where $k = 1$ to K is used to refer the relative type, g'_{ij} is the adjusted genotype of g_{ij} (that is down-weighted by twice the kinship coefficient between the i pair of proband and relative), ${g'}_i^{Burden} = \sum_{j=1}^{m} w_j g'_{ij}$, $d_i(R_k)$ is the indicator variable that equals to 0 if the FH relative k for proband i is missing and equals to 1 otherwise. Then the score $\tilde{S}_{R_k} = \left(\tilde{S}_{R_{k_1}}, \tilde{S}_{R_{k_2}}, \dots, \tilde{S}_{R_{k_m}}\right)$ with calibrated variance can be obtained,

$$\tilde{S}_{R_k} \sim MVN\ (0{,}0, \left(\frac{\tilde{V}^R}{\tilde{r}^R}\right)^{\frac{1}{2}} D \left(\frac{\tilde{V}^R}{\tilde{r}^R}\right)^{\frac{1}{2}}), (5)$$

where $\tilde{V}^R = \left(\tilde{V}_1^R, \tilde{V}_2^R, \dots, \tilde{V}_m^R\right)$ with $\tilde{V}_j^R = \frac{{S^2}_{R_{k_j}}}{\chi^2_{quantile}\left(1 - \tilde{p}_j\ \right)}$, and $\tilde{r}^R = \min\ (1, (w^T \tilde{V}^{R^{\frac{1}{2}}} D \tilde{V}^{R^{\frac{1}{2}}} w) / \tilde{V}_{Burden}^R\ )$ where $\tilde{V}_{Burden}^R$ is the adjusted variance of S_{Burden}^R. Using (3) and (5), we developed robust versions of the methods that incorporates FH for rare variant association analysis,

$$robustQ = \sum_{j=1}^{m}\left(w_j^2 S_{Pj}^2 + \sum_k 4\Omega_\mathrm{k}{}^2 w_j^2 S_{R_{kj}}^2\right)$$

$$robustQ - O = (1-\rho)\sum_{j=1}^{m}(w_j^2 S_{Pj}^2 + \sum_k 4\Omega_\mathrm{k}{}^2 w_j^2 S_{R_{kj}}^2) + \rho\sum_{j=1}^{m}(w_j\, S_{Pj} + \sum_k 2\Omega_\mathrm{k}\, w_j\, S_{R_{kj}})^2,$$

where the parameter ρ is estimated to minimize the p-value.

The GLMMs (1) and (5) that account for relatedness are required to estimate the probability of having the disease under the null hypothesis of no genetic effects for studies containing related samples in probands or relatives, when calculating the robust-famFHAT and robust-famFHAT-O (i.e., robust versions of famFHAT and famFHAT-O). The standard logistic models that only consider covariates can be used to fit the null model for unrelated samples to calculate robust-FHAT and robust-FHAT-O (i.e., robust versions of FHAT and FHAT-O).

4.3 Simulation Studies

4.3.1 Empirical Significance

The methods that do not adjust for a case-control imbalance can lead to a high inflation in type I error when the disease prevalence is low. To make a fair power comparison, we can estimate the test-specific empirical significance estimated from the type I error analysis to evaluate the power for methods with high inflation when case-control ratio is extremely unbalanced instead of the nominal significance threshold. Specifically, we can calculate the empirical significance α_e such that the empirical probability of P (p-value < α_e)= the nominal alpha α. The power for methods that have inflated type I error in unbalanced case-control scenarios can be evaluated at the proportion of p-value less than or equal to the various α_e levels, while the power for the robust methods that have reasonable type I error are evaluated using the α.

4.3.2 Simulation Analysis in Unrelated Samples

4.3.2.1 Simulation Design

We took the same approach to evaluate type I error and power described in Wang et al. [57] for unrelated probands and relatives. To evaluate the performance of the robust methods in instances where the case-control is unbalanced, we considered prevalence ranging from 1% to 50%. We assigned two haplotypes simulated from HapGen2 [54] at random to 5000 mothers and 5000 fathers and randomly passed down one haplotype without recombination to their offspring (i.e., probands). The number of variants in the simulated regions was ~ 30. The model used to simulated the continuous phenotypes for 5000 probands and their parents is:

$$\begin{pmatrix} Y^P \\ Y^M \\ Y^F \end{pmatrix} = 0.015 \begin{pmatrix} age^P \\ age^M \\ age^F \end{pmatrix} + 0.25 \begin{pmatrix} sex^P \\ sex^M \\ sex^F \end{pmatrix} + \begin{pmatrix} {G_{causal}}^P \\ {G_{causal}}^M \\ {G_{causal}}^F \end{pmatrix} \gamma + \varepsilon, \quad (6)$$

Where $\varepsilon \sim MVN(0, \Sigma)$; Y^P, Y^M and Y^F represent the phenotypes of probands, mothers and fathers, respectively; age^P, age^M, age^F are ages for age of probands, mothers and fathers, which were simulated based on the age distribution in the UK Biobank; sex^P is the

Bernoulli distributed variable with fixed prevalence of 56% females among probands; sex^M is set to female for mothers and sex^F is set to males for fathers. The error term ε is assumed to have a multivariate normal distribution of mean 0 and covariance $\Sigma = \delta_g^2\Phi + \delta_e^2 I_{3\times 3}$, where Φ is twice the kinship matrix. The age and sex explained ~ 22% and 14% of the total variance, respectively, which allowed us to have an increased prevalence in older female individuals. Moreover, we used the vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)$ to control for the magnitude and direction of genetic effects of the causal variants that are defined in ${G_{causal}}^P$, ${G_{causal}}^M$, and ${G_{causal}}^F$ for probands, mothers and fathers. γ_j is fixed to 0 for the type I error simulation, and is defined as

$$\gamma_j = \sqrt{\frac{c}{2MAF_j\left(1 - MAF_j\right)}} \quad (7)$$

for power simulation, where $c = \frac{R^2}{V^T DV}$ is a constant, R^2= proportion of variance explained by causal variants, V is the vector assigning the direction for the genetic effects, and D is the correlation matrix among causal variants. We fixed $R^2 = 2\%$ in our simulation when all causal variants have positive effects, and $R^2 = 5\%$ when half of the causal variants have positive effects and half of the causal variants have negative effects. Parental genotypes were only used to simulate their phenotypes for the power simulation, and we assumed that parental genotypes were missing in all analyses.

For both power and type I error analysis, the normal approximation was applied to convert the continuous phenotypes to a binary trait with pre-specified prevalence using a liability scale. We used the same beta function (with parameters of 1, 25) to upweight rare variants for all methods. The standard logistic regression was used to fit the null model in the simulated dataset only containing unrelated samples. We evaluated the type I error of the robust methods (robust-FHAT, robust-FHAT-O, robust-SKAT, and robust-SKAT-O,) and compared to the results of non-robust methods (FHAT, FHAT-O, SKAT, SKAT-O, Burden, and ACAT-V) for prevalence = 1%, 5%, 10%, and 20%. The power was assessed for all methods for different prevalence.

4.3.2.2 Type I Error Results

We generated 20 million replicates to evaluate the performance of robust-FHAT and robust-FHAT-O and compared the results with robust versions of SKAT and SKAT-O (robust-SKAT and robust-SKAT-O). We also compared with non-robust methods that incorporate FH but do not adjust for unbalanced case-control ratios (FHAT and FHAT-O), and non-robust methods that do not adjust for FH or unbalanced case-control ratios (SKAT, SKAT-O, and Burden). The analysis results for the unrelated samples can be found in **Table 4.1**. When prevalence = 10% and 20% in probands, the type I error of robust methods (robust-FHAT, robust-FHAT-O, robust-SKAT, robust-SKAT-O) is similar to that of non-robust methods at alpha = 2.5×10^{-4}, and much less inflated than non-robust methods at alpha = 2.5×10^{-6}. In addition, SKAT, SKAT-O, and Burden always have the highest inflation in scenarios with unbalanced case-controls, while the inflation is reduced by methods incorporating additional FH from relatives because FH contributes additional cases. Compared to the non-robust methods that do not adjust for imbalance, the inflation is greatly reduced using the robust methods especially for the lower alpha level and when the case-control ratios are extremely unbalanced (P=1%, 5%, 10%).

Table 4.1 Type I error rates of robust methods and non-robust methods in unrelated samples

Alpha	robust-SKAT	SKAT	robust-FHAT	FHAT	robust-SKAT-O	SKAT-O	robust-FHAT-O	FHAT-O	Burden	ACAT-V
Prevalence = 20%										
2.5×10^{-4}	1.1	1.3	1.0	1.0	1.5	1.9	1.5	1.6	1.1	1.5
2.5×10^{-6}	1.1	3.0	1.2	1.6	2.0	6.4	2.8	3.5	1.7	1.9
Prevalence = 10%										
2.5×10^{-4}	1.2	2.3	1.1	1.3	1.5	2.9	1.7	1.9	1.5	1.6
2.5×10^{-6}	1.3	9.0	1.4	3.0	1.9	16.1	3.7	6.6	3.9	2.4
Prevalence = 5%										
2.5×10^{-4}	1.3	4.1	1.3	1.8	1.6	4.9	1.9	2.5	2.2	2.0
2.5×10^{-6}	1.6	25.9	2.2	6.3	2.4	40.2	4.7	11.4	8.8	7.4
Prevalence = 1%										
2.5×10^{-4}	1.9	16.9	2.2	4.9	1.9	18.4	2.9	6.1	6.8	6.2
2.5×10^{-6}	1.6	238.5	3.3	35.0	1.9	318.4	6.9	55.2	65.2	122.2

The number in each cell represents the ratio of type I error and expected significance level (column 'Alpha'). Type I error was evaluated from the proportion of p-values less than or equal to corresponding to each alpha level. robust-FHAT, robust-FHAT-O, FHAT, FHAT-O analyzed 5000 probands and incorporated the family history information, while robust-SKAT, robust-SKAT-O, SKAT, SKAT-O, Burden test and ACAT-V only included probands. All methods used the same Wu weights with beta (MAF_j; 1, 25) or the comparable weights. The analyses were restricted to rare variants with MAF< 1%.

4.3.2.3 Power Results

A total of 1000 replicates were generated to evaluate the power for unrelated samples and related samples separately. The power was calculated by the proportion of p-values less than equals to the empirical alpha levels for those methods that have largely inflated type I error in the scenarios of extremely low prevalence. Through the adjustment from empirical alpha levels, we can compare the results fairly because the type I error can be controlled correctly for all methods.

The power was evaluated in various scenarios where we considered proportion of causal variants = 10%, 20%, 50%, 80% and 100%; prevalence= 1%, 5%, and 10%; all causal variants are risk-increasing or half of variants are risk-increasing and half are risk-decreasing. **Figure 4.1** shows the power results of robust methods and the non-robust methods applied in unrelated samples at alpha= 2.5×10^{-6} for prevalence = 5%. (See Appendix C.1 for the results evaluated at prevalence = 1% and 10%). In most scenarios, robust-FHAT and robust-FHAT-O have greater or similar power compared to FHAT and FHAT-O, and greater power than robust-SKAT and robust-SKAT-O and other standard methods that do not incorporate FH. However, when all variants have positive effects and the proportion of causal variants ≥ 50%, the robust methods have slightly lower power than their non-robust versions. When the prevalence is very low, the proposed methods that incorporate FH have greatly improved power, while other standard methods have low power due to the limited number of proband cases.

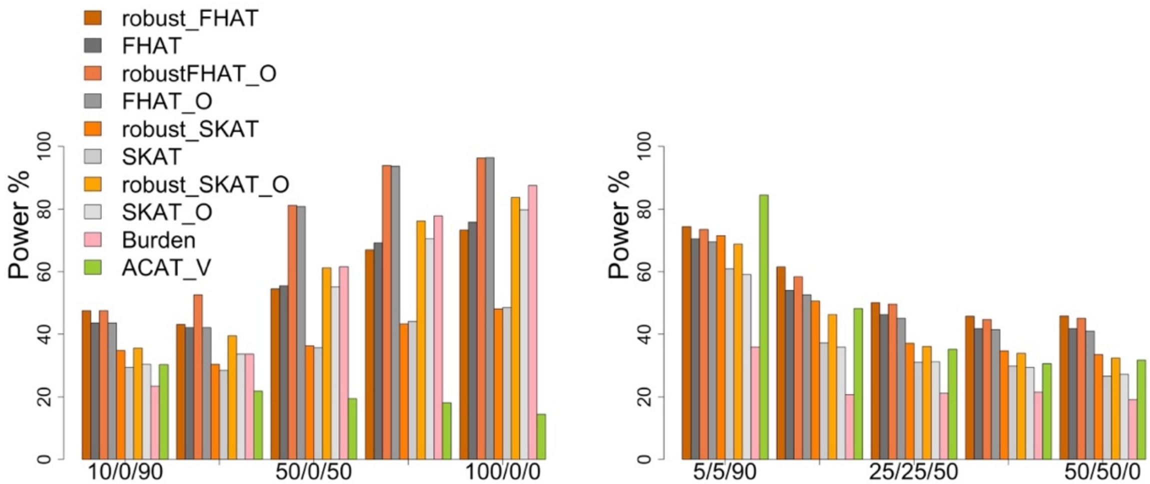

Figure 4.1 Empirical power of robust methods and non-robust methods in unrelated samples for prevalence = 5%

In each plot, the x axis in the format of +/-/0 indicates the proportion of variants with positive, negative and no effects. Each bar shows the empirical power evaluated as the proportion of p-values less than or equal to alpha= 2.5 x 10^{-6}. The empirical alpha level was used for the methods with inflated type I error to evaluate the power. The analyses were restricted to rare variants with MAF < 1%. All methods used the same Wu weights with beta (MAF_j; 1, 25). The analyses were restricted to rare variants with MAF < 1%.

4.3.3 Simulation Analysis in Related Samples

4.3.3.1 Simulation Design

The second simulation analysis investigated the performance of proposed methods accounting for both relatedness and unbalanced case-control designs for low prevalence in family studies. We simulated 400 large families with 18 members based on the pedigree shown in **Figure 3.2**. All founders were randomly assigned two simulated haplotypes and pass down one of the haplotype to their kids. We simulated the phenotypes for all 7200 samples using the similar model (6) for simulating 3 family members (mother, father and proband), but to simulate 18 family samples, where ε assumed to follow a multivariate normal distribution with mean 0 and covariance matrix $\Sigma = 0.4 * \Phi + 0.6 * I_{3\times3}$ where Φ

is twice the kinship matrix among 18 individuals within a family. The elements in γ are defined in the same way as in (7) to represent the genetic effects of causal variants. The simulated continuous phenotypes were converted to a binary outcome with different pre-specified prevalence using the normal approximation. The proportion of variance explained by causal variants R^2 was fixed to 2% for the scenarios where all genetic effects have the positive directions, and 3% for the scenarios where ½ of the causal variants have the positive effects and ½ have negative effects. We analyzed the regions with 30 variants.

We randomly selected 9 family members as the probands and kept the remaining individuals as relatives for each single family. In each replicate, we followed this process and simulated the dataset containing 3600 probands with available FH from 3600 relatives. The relatives were omitted if they were not related to any probands in the randomly simulated dataset. In the analysis, we used both phenotype and genotype data for probands, and only used the phenotype data for relatives (i.e. FH) to calculate robust-famFHAT, robust-famFHAT-O, famFHAT and famFHAT-O, while the methods that ignore FH only used the probands' data.

4.3.3.2 Type I Error Results

The performance of the robust methods for related samples was also investigated (**Table 4.2**). We generated 10,000 replicates to evaluate the type I error at alpha level as low as 0.005 to reduce the computational cost. The results showed that robust-famFHAT and robust-famFHAT-O control the type I error well in family samples while the non-robust methods suffer from substantial inflated type I error when the case-control ratios are unbalanced. We also observed that the robust methods have deflated type I error compared to the methods without adjustment for unbalanced case-control ratios for higher disease prevalence (P= 20%, 50%), which was consistent to what has been observed in previous investigations. [60,61,62] When the prevalence is high (P= 50%), famFHAT and famFHAT-O have slightly higher inflation, which might be caused by the residual correlation among scores of probands and relatives since we only conditioned on data from the closest probands in the relatives' analysis. The robust-famFHAT and robust-famFHAT-O based on the GLMMs can control the type I error rate well in the presence of family samples while greatly reducing the type I error inflation for low prevalence.

Table 4.2 Type I error rates of robust methods and non-robust methods in related samples

Alpha	robust-famSKAT	famSKAT	robust-famFHAT	famFHAT	robust-famSKAT-O	famSKAT-O	robust-famFHAT-O	famFHAT-O
Prevalence = 50%								
0.1	0.6	1.0	0.8	1.2	0.7	1.0	0.8	1.1
0.05	0.6	1.1	0.6	2.1	0.7	1.5	0.7	1.6
0.01	0.4	0.9	0.7	1.2	0.5	1.0	0.7	1.2
0.005	0.3	0.7	0.6	1.3	0.4	0.9	0.6	1.4
Prevalence = 20%								
0.1	0.7	1.0	0.8	1.1	0.8	0.9	0.8	1.0
0.05	0.7	1.0	0.7	1.1	0.8	1.0	0.9	1.0
0.01	0.5	0.8	0.6	1.0	0.6	1.1	0.8	1.1
0.005	0.5	0.8	0.5	1.0	0.7	1.0	0.8	1.3
Prevalence = 10%								
0.1	0.8	1.0	0.9	1.1	0.9	1.0	0.9	1.0
0.05	0.8	1.1	0.9	1.1	0.8	1.0	0.9	1.1
0.01	0.8	1.2	0.9	1.3	0.9	1.3	1.0	1.4
0.005	0.8	1.3	1.0	1.5	0.9	1.3	1.1	1.6
Prevalence = 5%								
0.1	0.8	1.0	0.9	1.1	0.9	1.0	1.0	1.0
0.05	0.9	1.1	1.0	1.2	0.9	1.0	1.0	1.1
0.01	1.1	1.5	1.1	1.5	1.1	1.4	1.1	1.4
0.005	1.0	1.9	1.0	1.9	0.9	1.9	1.1	1.7

The number in each cell represents the ratio of type I error and expected significance level (column 'Alpha'). Type I error was evaluated from the proportion of p-values less than or equal to corresponding to each alpha level. All methods used the same Wu weights with beta (MAF_j; 1, 25). The analyses were restricted to rare variants with MAF< 1%.

4.3.3.3 Power Results

The results in **Figure 4.2** summarizing the power for related samples show the similar patterns as in the power simulation for unrelated samples. The power was evaluated at alpha= 0.005 reduce the computational cost, where we adjusted the test-specific empirical alpha estimated from type I error results to evaluate the non-robust methods. Overall, the robust methods perform similarly or better than the unadjusted methods in a wide range of scenarios. When the proportion of causal variants is larger than or equals to 50%, the robust versions have slightly lower power than their non-robust versions when all variants in the region are risk-increasing.

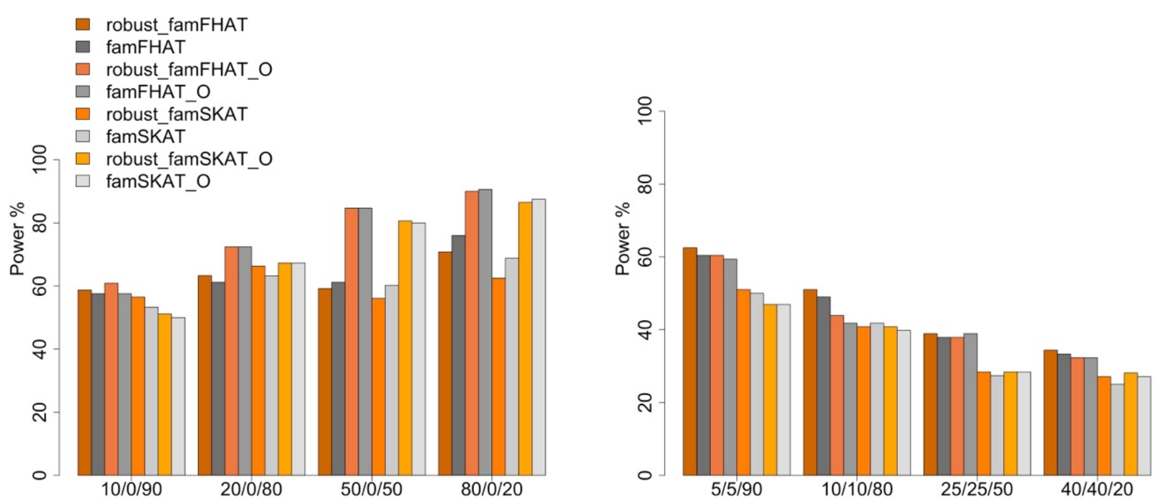

Figure 4.2 Empirical power of robust methods and non-robust methods in related samples at prevalence = 5%

In each plot, the x axis in the format of +/-/0 indicates the proportion of variants with positive, negative and no effects. Each bar shows the empirical power evaluated as the proportion of p-values less than or equal to alpha= 0.005. The empirical alpha level was used for the methods with inflated type I error to evaluate the power. All methods used the same Wu weights with beta (MAF_j; 1, 25). The analyses were restricted to rare variants with MAF < 1%.

4.4 Applications

4.4.1 Analysis of Whole Exome Sequencing Data in the UK Biobank

We first applied the robusFHAT and robustFHAT-O to the UK Biobank, the same dataset we used in the Chapter 2. We analyzed the same exome sequence data that processed using the functional equivalence (FE) pipeline and the same quality control (QC) procedure as in Chapter 2. Because the robust-famFHAT and robust-famFHAT-O were developed using the GLMM adopted in GMMAT [5], it is not computationally feasible for the large-scale analysis. We restricted the analysis to the unrelated white population. We randomly omitted one proband in each related pair with 1st, 2nd, or 3rd degree of relationships to select unrelated samples. The analysis was performed to analyze all cause dementia in white participants who are self-identified as White, British, Irish, and any other white groups. Missing genotypes in probands were imputed using the mean genotypes to calculate the scores for probands and relatives. The rare coding variants with minor allele frequency (MAF) < 1% were selected in the analysis. Following the same analysis scheme in the application in Chapter 2, we selected age, sex, PC1-PC5 and PC11 as the covariates for all cause dementia analysis. The details for the computation and inclusion of principle components can be found in [57].

We conducted exome-wide analyses using the robust-FHAT and robust-FHAT-O for all cause dementia, a trait with low prevalence in UK Biobank participants (P= 0.02% in probands, P= 8.2% in mothers, P= 4.4% in fathers). The goal of this analysis was to compare the results from the robust methods, the most appropriate approaches given the low prevalence, to those obtained using non-robust versions of FHAT and FHAT-O. A total of ~18K genes were tested based on 129,670 unrelated white individuals who pass QC and have available all dementia status and parental history.

The genomic control inflation factor $\lambda_{\text{robust}-\text{FHAT}}$ of robust-FHAT was 1.11, which was slightly reduced compared to that was obtained using the FHAT (λ_{FHAT} =1.13). The $\lambda_{\text{robust}-\text{FHAT}-\text{O}}$ of robust-FHAT-O ($\lambda_{\text{robust}-\text{FHAT}}$=1.04) was also reduced compared to the FHAT-O ($\lambda_{\text{FHAT}-\text{O}}$ =1.06). Interestingly, the robust-FHAT and robust-FHAT-O identified the same genes as those found using non-robust FHAT and FHAT-O in our prior analysis with a suggestive threshold of 5.6x10^{-5}. The top associations are reported in **Table 4.3**. Among those genes that passed the exome-wide based significance threshold, the p-values calculated using the robust-FHAT were generally larger (less significant) than those calculated using FHAT, and the p-values calculated using the robust-FHAT-O were larger (less significant) than those obtained using FHAT-O, but those p-values estimated using the robust methods were still below the significance cut-off. Those results confirmed our previous findings with non-robust FHAT and FHAT-O. Robust methods accounting for the low prevalence in all cause dementia in the UK Biobank provided more accurate p-values for the detected genes.

Table 4.3 Exome-wide analysis for all cause dementia in the UK Biobank

Gene	#variants	cumMAC	robust-FHAT (p-value)	robust-FHATO (p-value)	FHAT (p-value)	FHAT-O (p-value)
TREM2	45	4559	6.6 x10^{-6}	1.2 x10^{-8}	5.2 x10^{-6}	4.1 x10^{-9}
PVR	75	2068	1.9 x10^{-5}	2.7 x10^{-5}	1.2 x10^{-5}	1.8 x10^{-5}
EFCAB3	60	2579	4.6 x10^{-5}	5.0 x10^{-5}	4.0 x10^{-5}	4.2 x10^{-5}
EMSY	158	1543	7.2 x10^{-5}	3.9 x10^{-5}	4.4 x10^{-5}	2.7 x10^{-5}
BCL3	65	1157	8.8 x10^{-5}	7.8 x10^{-5}	6.8 x10^{-5}	5.9 x10^{-5}
KLC3	177	4174	5.1 x10^{-4}	1.6 x10^{-5}	4.9 x10^{-4}	1.3 x10^{-5}
ABCA7	487	12178	2.9 x10^{-3}	4.2 x10^{-5}	2.9 x10^{-3}	4.1 x10^{-5}

The exome-wide significance threshold is $\frac{0.05}{18,000} = 2.8 \times 10^{-6}$ the suggestive exome-wide significance threshold is $\frac{1}{18,000} = 5.6 \times 10^{-5}$. cumMAC is the cumulative minor allele frequency in the region. #variants is the total number of variants in the gene.

4.4.2 Analysis of Exome Chip Data in the Framingham Heart Study

As an illustration of robust-famFHAT and robust-famFHAT-O, we applied these methods to re-investigate the gene-disease associations using the same dataset as in Chapter 3. We followed the same QC process to select the coding rare variants with call rate > 95% and MAF < 1%. We used the same models for AD, dementia, and type 2 diabetes (T2D) as in Chapter 4. While there were genes with very similar p-values estimated using robust methods and non-robust methods, we found that *PYGM* was not associated with AD anymore with a suggestive significance threshold using the robust methods. However, there were scenarios where the robust-famFHAT yielded the smaller p-values compared to the famFHAT version such as *ZSCAN18* and *KIAA0368* for AD and dementia.

As shown in the **Table 4.5**, the p-values for genes became insignificant using robust-famFHAT and robust-famFHAT-O, compared to the results of the non-robust methods (famFHAT and famFHAT-O) for the T2D analysis.

Table 4.4 Exome chip analysis of AD and dementia

Gene	#variants	cumMAC	robust-famFHAT (p-value)	robust-famFHATO (p-value)	famFHAT (p-value)	famFHAT-O (p-value)
AD						
CYP26B1	3	20	5.1×10^{-4}	4.8×10^{-4}	8.4×10^{-5}	8.8×10^{-5}
CCR5	7	70	4.9×10^{-4}	2.3×10^{-4}	4.4×10^{-4}	9.9×10^{-5}
ODZ3	13	92	2.3×10^{-4}	3.2×10^{-4}	9.7×10^{-5}	1.2×10^{-4}
KIAA0368	9	108	1.6×10^{-4}	6.2×10^{-5}	2.2×10^{-4}	3.6×10^{-5}
PYGM	14	133	4.6×10^{-4}	3.8×10^{-4}	8.4×10^{-5}	7.4×10^{-5}
ZSCAN18	2	38	4.4×10^{-5}	4.1×10^{-5}	7.6×10^{-5}	7.0×10^{-5}
Dementia						
KIAA0368	9	108	5.7×10^{-5}	2.7×10^{-5}	7.1×10^{-5}	2.2×10^{-5}

cumMAC is the cumulative minor allele counts in probands for the gene we tested. #variants is the total number of variants tested in the gene. The suggestive significance threshold is $\frac{1}{6{,}831} = 1.5 \times 10^{-4}$.

Table 4.5 Exome chip analysis of T2D

Gene	#variants	cumMAC	robust-famFHAT (p-value)	robust-famFHAT-O (p-value)	famFHAT (p-value)	famFHAT-O (p-value)
PDE8A	4	22	1.9×10^{-2}	1.1×10^{-3}	4.2×10^{-3}	1.1×10^{-4}
MAP3K3	3	29	1.8×10^{-3}	2.2×10^{-4}	8.7×10^{-4}	1.2×10^{-4}

cumMAC is the cumulative minor allele counts in probands for the gene we tested. #variants is the total number of variants tested in the gene. The suggestive significance threshold is $\frac{1}{8{,}218} = 1.2 \times 10^{-4}$.

4.5 Discussion

In this project, we proposed robust versions of our previously developed methods that allow the incorporation of FH for rare variant association analysis with enhanced statistical evidence to detect complex disease in related samples and unrelated samples. The robust versions were proposed based on cutting-edge approaches of SPA and ER. Because SPA is an asymptotic-based approach and does not perform well for p-value calculation when a variant is extremely rare, the ER method is used for very rare variants with minor allele count (MAC) ≤ 10. In the robust methods, the variance for the single score statistic is calibrated through the p-values calculated from SPA or EA when case-control ratios are extremely unbalanced, which successfully addresses the inflation of type I error that occurs in studies with low disease prevalence.

Two simulation analyses were performed to evaluate the type I error and power for our newly proposed robust methods in unrelated samples and related samples. We concluded that the robust-FHAT and robust-FHAT-O can control type I error well in unrelated samples when case-control ratios are moderately or extremely high. The robust-famFHAT and robust-famFHAT-O were developed to correct the type I error rate of low prevalent binary traits while adjusting for relatedness among samples in family studies. The correct type I error using robust methods was demonstrated in simulation analyses, and we also showed that non-robust methods have large inflation in type I error when the prevalence is low. However, the slight inflation was still observed in the robust methods for extremely low prevalence (prevalence = 10%), the genomic control can be used to improve the type

I error further as suggested by Zhou et al. [62] Because the non-robust methods had severe inflation in type I error when the number of case and control samples was extremely unbalanced, the empirical significance was used to evaluate their power. The power results showed that the robust methods perform better or equally compared to their non-robust versions that do not account for case-control ratios in most cases. The robust-FHAT and robust-FHAT-O maintain a high statistical power after incorporating FH compared to robust-SKAT and robust-SKAT-O while adjusting for the case-control imbalance. With the same improvement in power and type I error as robust-FHAT and robust-FHAT-O, robust-famFHAT and robust-famFHAT-O can also account for correlation in family samples.

We applied the robust-FHAT and robust-FHAT-O to analyze all dementia using the whole exome sequencing data for the unrelated white UK Biobank samples. Our results confirmed the previous findings obtained using non-robust FHAT and FHAT-O, where more conservative p-values and lower inflation factors were observed in robust methods. The robust-famFHAT and robust-famFHAT-O were developed based on the same GLMM as in SMMAT, [5] which is not computationally feasible for the large cohort and thus prevented us from applying robust-famFHAT and robust-famFHAT-O to related samples in the UK Biobank. In the future, we will take similar optimization strategies adopted in BOLT-LMM [41] and SAIGE, to make the methods computationally practical for large data. Instead, we applied robust-famFHAT and robust-famFHAT-O to the FHS to conduct same analysis as in Chapter 3.

In summary, we proposed the robust methods that can correctly adjust type I error for binary traits with unbalanced case-control ratios, while greatly increasing the association power by exploiting FH when available. The robust-FHAT and robust-FHAT-O require all probands to be unrelated, while the robust-famFHAT and robust-famFHAT-O can be applied to family data with related individuals. The insufficient cases in younger probands or missing genotypes in older patients limit the power of many cohorts or biobanks for the genetic associations, especially for the study of late-onset disease (i.e., AD). Our method incorporating FH is the most effective way to overcome those limitations. With the accurate results provided in the robust methods, our methods will significantly contribute to the identification of trait associated with rare variants.

Chapter 5 Summary and Future Work

5.1 Summary

Genome-wide association studies (GWAS) are designed to assess the associations between common diseases and common variants, defined as variants with minor allele frequency (MAF) greater than 1% or 5%. Thousands of variants associated with traits haven been successfully identified by GWAS, hence contributing to the understanding of genetic etiology of diseases. Despite this big achievement, the common variants identified by GWAS only explain a small portion of the total heritability. Unexplained heritability may be due to causal variants which remain undetected due to limited statistical power that typically depends on sample size and variants frequency. To address these challenges, we develop methods that increase the sample size from incorporating additional family history (FH) data to boost the statistical power for rare variant associations. In Chapter 2, we propose novel methods that enable the incorporation of FH based on a variance component framework for rare variant analysis: family history aggregation unit-based test (FHAT) and optimal FHAT (FHAT-O). These methods are developed using generalized linear models (GLMs) that can accommodate both binary and continuous traits with a focus on unrelated samples. In Chapter 3, we develop family-based methods (famFHAT and famFHAT-O) accounting for sample relatedness through generalized linear mixed models (GLMMs) to avoid false-positive results in studies containing related samples. We also present the strategy to combine FH from multiple relatives with various types of relationship as would be available in complex study cohorts such as the Framingham heart study (FHS). In

Chapter 4, we incorporate saddle point approximation (SPA) and efficient resampling (ER) to calibrate the variance of score statistic, thus providing more accurate estimates of statistical significance when studies have extremely unbalanced case-control ratios.

With rapid advances in whole genome sequencing and whole exome sequencing technologies, the proposed methods provide a cost-effective way to overcome power limitations in studies with limited sample sizes and low disease cases, by exploiting additional data from FH. The findings enabled by these powerful methods will help further characterize the underlying disease mechanism.

5.2 Future Work

5.2.1 Accuracy of Family History

Although we demonstrate the invaluable contribution of FH in genetic analysis, concerns remain about the accuracy of the reported FH. It's important to leverage high accurate information reported in FH to make valid statement about the association between genetic markers and traits. One way to access the accuracy of FH is to calculate the correlation of reported FH between sibling pairs. Hujoel et al. [26] calculated the correlation of self-reported sibling history using the analysis restricted to all cases or all controls, and they accounted for the FH accuracy by down-weighting FH information. We may modify our methods in a similar way to adjust for FH accuracy in the future work.

5.2.2 Extensions of FHAT-O and famFHAT-O

We developed FHAT-O and famFHAT-O based on the same framework in optimal sequence kernel association test (SKAT-O) to maintain robust power in various scenarios regardless of direction of genetic effects or proportion of causal variants. The SMMAT-E approach that was developed based on Mixed effects Score Test (MiST) has shown to be more powerful and computationally feasible compare to SKAT-O. Therefore, future efforts to improve FHAT-O and famFHAT-O by incorporating SMMAT-E will be investigated.

5.2.3 Computational Feasibility for Large-Scale Samples

The same GLMMs in the Generalized linear Mixed Model Association Tests (GMMAT) are used in our proposed methods to fit the null model to account for family correlation. However, those methods are not computationally feasible for large-scale dataset. We plan to incorporate the state-of-the-art optimization strategies to lower the computational time and memory usage.

5.2.4 Gene-environmental Iteration

The roles of genetic and environmental factors in the development of complex diseases have been greatly recognized. Investigation of the genetic and environmental interaction can bring important insights on biological mechanism underlying the etiology of complex diseases. [24,44] Several methods have been proposed to include the genetic and environmental interaction for rare variant analysis. [7,9] We will expand the proposed methods to detect genetic and environmental interactions for rare variant analysis to correctly reveal their effect on the etiology, while maintaining the optimal power.

APPENDIX A

Supplementary Material for Chapter 2

A.1: Additional Type I Error Analysis

The Type I error for FHAT and FHAT-O was evaluated at exome-wide significance and compared to other methods. Here we did not compare to SKAT-LTFH and SKATO-LTFH due to a high CUP time consumption. We considered the disease prevalence = 10%, 20% 30% and 50%, and we generated 20 million replicates for each scenario. With p being fixed as ~ 10%, 20%, 30% and 50%, we had the prevalence ~ 21%, 35%, 41%, and 69% in mothers, and the prevalence ~ 16%, 28%, 49%, and 62% in fathers. We first simulated Y^P for 5000 probands and their mothers Y^M and fathers Y^F using the same simulation model presented in the Simulation to Evaluate Type I Error section.

Table A.1 Type I error rates of FHAT, FHAT-O, SKAT, SKAT-O, Burden and ACAT-V

Prevalence	FHAT	SKAT	FHAT-O	SKAT-O	Burden	ACAT-V
Alpha= 2.5 x 10^{-4}						
P=10%	1.4	2.4	1.6	2.7	1.2	1.3
P=20%	0.9	1.3	1.2	1.6	1.0	1.2
P=30%	0.9	1.0	1.2	1.2	1.0	1.1
P=50%	0.9	0.8	1.1	1.0	1.0	0.9
Alpha= 2.5 x 10^{-6}						
P=10%	3.2	8.6	3.8	10.9	2.0	2.1
P=20%	1.3	2.7	2.0	4.0	1.6	1.2
P=30%	0.8	1.5	1.4	2.4	1.3	1.3
P=50%	0.6	0.5	1.0	1.0	1.0	0.9

The number in each cell represents the ratio of type I error and expected significance level. Type I error was evaluated from the proportion of p-values less than or equal to corresponding 2.5 x 10^{-4} and 2.5 x 10^{-6} using 20 million simulation replicates. The x axis is the disease prevalence of 5%, 10%, and 20%. The total sample size of probands were 5000. FHAT FHAT-O, SKAT, SKAT-O and Burden test all used the same Wu weights with beta (MAF_j; 1, 25). ACAT-V used the weight of $w_{j,ACAT-V} = w_{j,SKAT} \times \sqrt{MAF_j\,(1 - MAF_j)}$ to make results comparable. FHAT and FHAT-O analyzed probands and incorporated the family history information, while SKAT, SKAT-O, Burden test and ACAT-V only included probands.

A.2: Additional Power Analysis

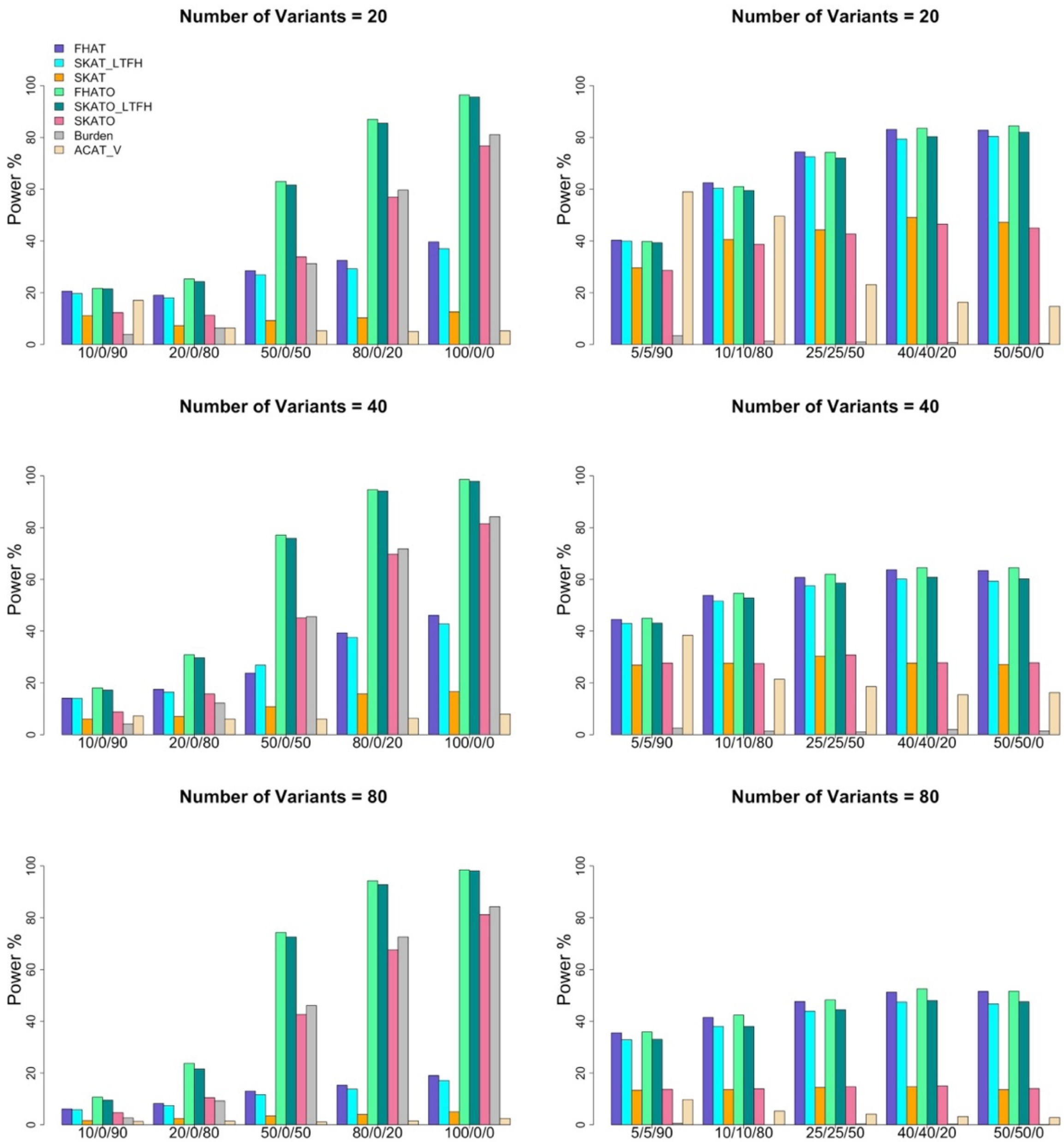

Figure A.1 Empirical power of FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, Burden test and ACAT-V estimated at alpha= 2.5 x 10^{-5} for prevalence = 50%

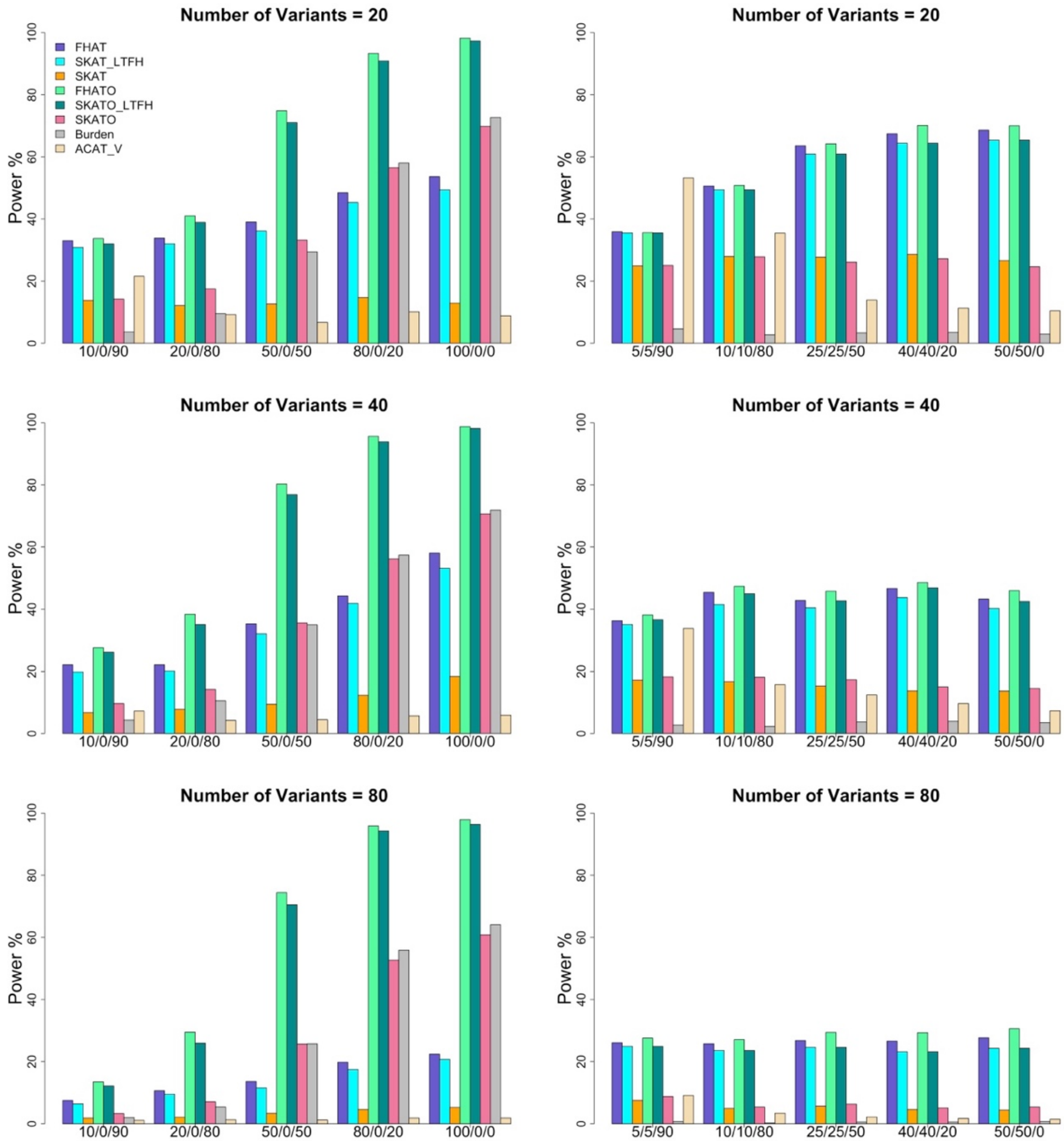

Figure A.2 Empirical power of FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, Burden test and ACAT-V estimated at alpha= 2.5 x 10^{-6} for prevalence = 20%

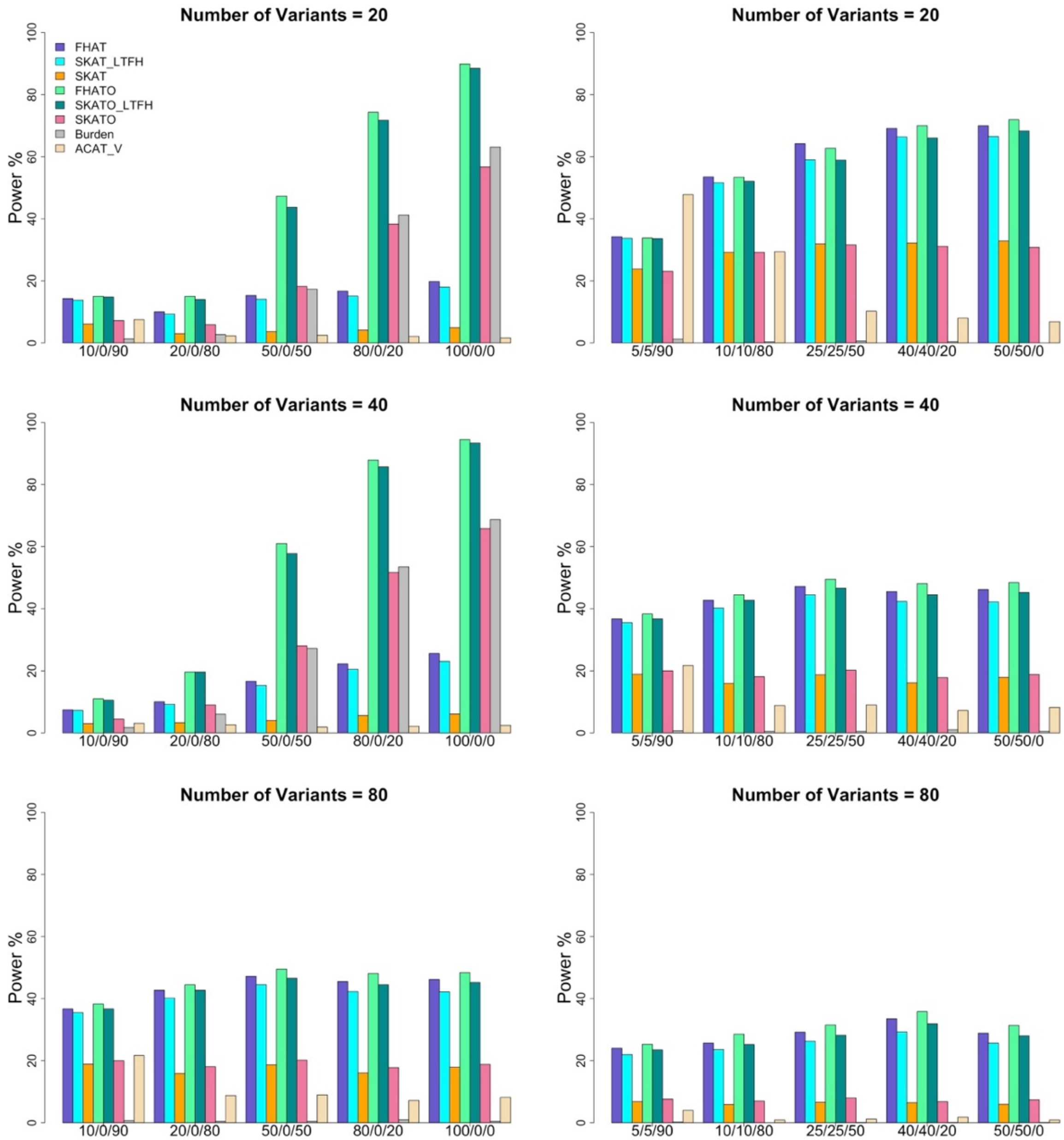

Figure A.3 Empirical power of FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, Burden test and ACAT-V estimated at alpha= 2.5 x 10^{-6} for prevalence = 50%

A.3: UK Biobank Association Analysis between PCs and Disease of Interest

Before selecting the final model for all cause dementia and hypertension analysis, we first tested the significance between diseases (all cause dementia and hypertension) and each of the 20 PCs in parents and probands separately, and meta-analyzed the results. We used the probands' PCs as the proxy-PCs for relatives. We selected the top 5 PCs and significant PCs (P< $\frac{0.05}{20} = 2.5\times10^{-3}$) to adjust for population structure in the analysis.

Table A.2 P-values for the association analysis between PCs and diseases in the UK Biobank

	All cause dementia	**Hypertension**
PC1	0.01	0.71
PC2	0.53	0.37
PC3	0.51	0.4
PC4	0.31	4.0×10^{-3}
PC5	2.0×10^{-3}	0.024
PC6	0.96	0.02
PC7	0.26	0.11
PC8	0.037	9.3×10^{-5}
PC9	0.19	0.045
PC10	0.27	0.89
PC11	6.2×10^{-4}	0.16
PC12	0.56	0.11
PC13	0.34	0.53
PC14	0.2	6.8×10^{-6}
PC15	0.78	0.14
PC16	6.6×10^{-3}	0.052
PC17	0.028	0.2
PC18	0.24	0.81
PC19	0.78	0.82
PC20	0.16	3.7×10^{-3}
The top 5 PCs and significant PCs were selected in the testing models. Using the significance threshold of 2.5×10^{-3} for testing 20 PCs, we adjusted PC1-5, and PC11 were adjusted in all cause dementia model, and PC1-5, PC8 and PC14 were adjusted in hypertension model.		

APPENDIX B

Supplementary Material for Chapter 3

Appendix B.1: The Score Statistic in the Generalized Linear Mixed Model

We use a Generalized Linear Mixed Model (GLMM) to account for correlation between observations through a random effect $\delta \sim N(0, \sigma_G^2 \Phi)$:

$$g\big(E(Y^P | G^P, X^P, \delta)\big) = X^P \alpha_P + G^P \beta_P + \delta ,$$

where $g(\cdot)$ is the link function that connects the phenotype mean Y^P with the covariate matrix X^P, the genotype matrix G^P and the random effect δ. We assume the $n \times 1$ random effect vector follows the distribution of $N(0, \sigma_{G_P}^2 \Phi_\mathrm{P})$, where Φ_P is the $n \times n$ matrix containing twice the kinship matrix, and $\sigma_{G_P}^2$ is the parameter of variance component.

To estimate the parameters for the GLMM at the null hypothesis, we define the working vector

$$Y^{P^*} = \mathcal{R}(Y^P - \mu_P) + \omega,$$

where $\omega_i = g(\mu_{P_i})$, $\mathcal{R} = diag\{\left(\frac{\partial \mu_{P_i}}{\partial \omega_i}\right)^{-1}\} = diag\{\left(\frac{\partial g^{-1}(\omega)}{\partial \omega}\right)^{-1}\}$. We also denote $V_P = diag\{Var(Y_i^P)/[g'(\omega_i)]^2\}^{-1} = diag\left\{Var(Y_i^P)/\left(\frac{\partial \mu_{P_i}}{\partial \omega_i}\right)^2\right\}^{-1}$. Then the variance-covariance matrix $\hat{\Sigma}_P = \hat{\sigma}_G^2 \Phi + {V_P}^{-1}$.

- For the continuous trait, one can show that $\mathcal{R} = I, Y^{P^*} = \omega, {V_P}^{-1} = \phi I$, then $\hat{\Sigma}_P = \hat{\sigma}_G^2 \Phi + \hat{\phi} I$.

- For binary trait, $\omega = log \frac{\mu_P}{1-\mu_P}$, then $logit^{-1}(\omega) = \mu_P$, then we take the derivative

  of μ_P and get

$$\frac{\partial \mu_P}{\partial \omega} = \left(\frac{e^{\omega}}{(1+e^{\omega})^2}\right) = \left(\frac{e^{\omega}}{1+e^{\omega}}\right)\left(1-\frac{e^{\omega}}{1+e^{\omega}}\right) = \mu_P(1-\mu_P),$$

  so that we have $V_P = diag\{\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n)\}$, and thus,

  $\hat{\Sigma}_P = \hat{\sigma}_G^2 \Phi + diag\{1/(\hat{\mu}_{Pi}(1-\hat{\mu}_{Pi}))\}$, where $\hat{\phi} = 1$.

APPENDIX C

Supplementary Material for Chapter 4

Appendix C.1: Additional Power Analysis

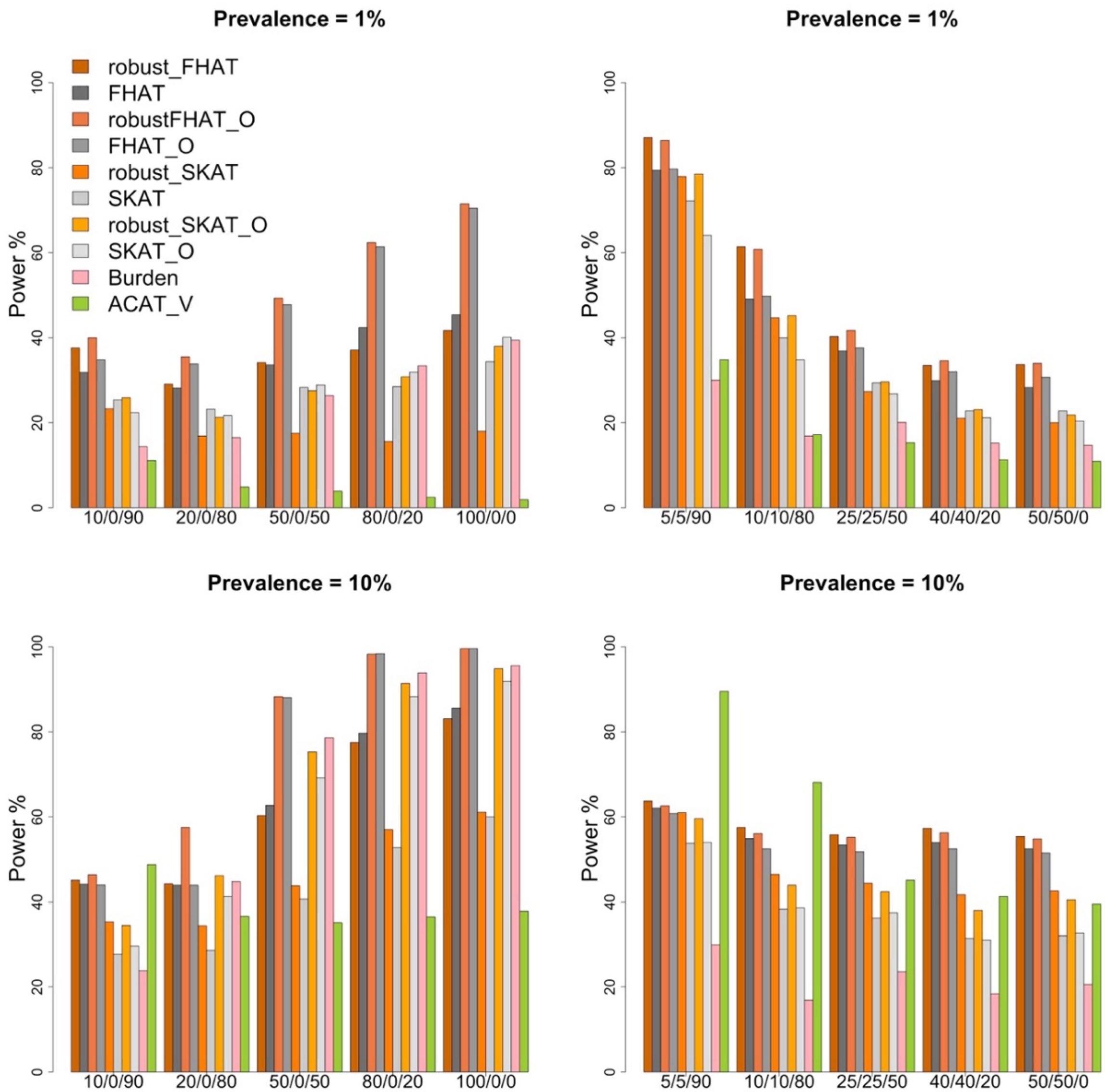

Figure C.1: Empirical power of robust methods and non-robust methods in unrelated samples for prevalence = 1% and 10%

BIBLIOGRAPHY

1. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101 (2002).
2. Bachman, D.L. *et al.* Prevalence of Dementia and Probable Senile Dementia of the Alzhermer type in the Framingham-Study. *Neurology* **42**, 115–119 (1992).
3. Bellenguez, C. *et al.* New insights on the genetic etiology of Alzheimer's and related dementia. *medRxiv*, DOI: 10.1101/2020.10.01.20200659 (2020).
4. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics* **40**, 695–701 (2008).
5. Chen, H. *et al.* Efficient Variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *American Journal of Human Genetics* **104**, 260–274 (2019).
6. Chen, H., Meigs, J.B. & Dupuis, J. sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology* **37**, 196–204 (2013).
7. Chen, H., Meigs, J.B. & Dupuis, J. Incorporating gene-environment interaction in testing for association with rare genetic variants. *Human Heredity* **78**, 81–90 (2014).
8. Chen, H. *et al.* Control for Population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *American Journal of Human Genetics* **98**, 653–666 (2016).

9. Coombes, B.J., Basu, S. & McGue, M. A linear mixed model framework for gene-based gene-environment interaction tests in twin studies. *Genetic Epidemiology* **42**, 648–663 (2018).

10. Daniels, H.E. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, **25**, 631–650 (1954).

11. Davies, R.B. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Applied Statistics*, **29**, 323–333 (1980).

12. Dawber, T.R., Kannel, W.B. & Lyell, L.P. An approach to longitudinal studies in a community: the Framingham Study. *Annals of the New York Academy of Sciences* **107**, 539–556 (1963).

13. Dey, R., Schmidt, E.M., Abecasis, G.R. & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *American Journal of Human Genetics* **101**, 37–49 (2017).

14. Dolgin, E. Personalized investigation. *Nature Medicine* **16**, 953–955 (2010).

15. Ehret, G.B. *et al.* The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nature Genetics* **48**, 1171–1184 (2016).

16. Eichler, E.E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews. Genetics* **11**, 446–450 (2010).

17. Eriksson, N. *et al.* Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genetics* **6**, e1000993 (2010).

18. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry* **63**, 168–174 (2006).

19. Ghosh, A. *et al.* Leveraging Family History in Population-Based Case-Control Association Studies. *Genetic Epidemiology* **38**, 114–122 (2014).

20. Ghosh, A. *et al.* Assessing Disease Risk in Genome-wide Association Studies Using Family History. *Epidemiology* **23**, 616–622 (2012).

21. Gim, J. *et al.* Improving Disease Prediction by Incorporating Family Disease History in Risk Prediction Models with Large-Scale Genetic Data. *Genetics* **207**, 1147–1155 (2017).

22. Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nature Genetics* **51**, 51–62 (2019).

23. Gratuze, M., Leyns, C.E.G. & Holtzman, D.M. New insights into the role of TREM2 in Alzheimer's disease. *Molecular Neurodegeneration* **13**, 66–66 (2018).

24. Hamza, T.H. *et al.* Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. *PLoS genetics* **7**, e1002237–e1002237 (2011).

25. Hollingworth, P. *et al.* Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nature Genetics* **43**, 429–435 (2011).

26. Hujoel, M.L.A., Gazal, S., Loh, P.-R., Patterson, N. & Price, A.L. Liability threshold modeling of case–control status and family history of disease increases association power. *Nature Genetics* **52**, 541–547 (2020).

27. Jansen, I.E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics* **51**, 404–413 (2019).

28. Jiang, D. & McPeek, M.S. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genetic Epidemiology* **38**, 10–20 (2014).

29. Kannel, W.B., Feinleib, M., McNamara, P.M., Garrison, R.J. & Castelli, W.P. An investigation of coronary heart disease in families. The Framingham offspring study. *American Journal of Epidemiology* **185**, 1093–1102 (2017).

30. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310–315 (2014).

31. Kuonen, D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929–935 (1999).

32. Ladouceur, M., Dastani, Z., Aulchenko, Y.S., Greenwood, C.M.T. & Richards, J.B. The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS Genetics* **8**, e1002496 (2012).

33. Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X.H. Rare-Variant association analysis: study sesigns and statistical tests. *American Journal of Human Genetics* **95**, 5–23 (2014).

34. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics* **91**, 224–237 (2012).

35. Lee, S., Fuchsberger, C., Kim, S. & Scott, L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. *Biostatistics* **17**, 1–15 (2016).

36. Lee, S., Teslovich, T.M., Boehnke, M. & Lin, X.H. General framework for meta-analysis of rare variants in sequencing association studies. *American Journal of Human Genetics* **93**, 42–53 (2013).

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

38. Li, B.S. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics* **83**, 311–321 (2008).

39. Liu, J.Z., Erlich, Y. & Pickrell, J.K. Case-control association mapping by proxy using family history of disease. *Nature Genetics* **49**, 325–331 (2017).

40. Liu, Y.W. *et al.* ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *American Journal of Human Genetics* **104**, 410–421 (2019).

41. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015).

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

43. Marioni, R.E. *et al.* GWAS on family history of Alzheimer's disease. *Translational Psychiatry* **8**, 161 (2019).

44. Matsui, T. & Ehrenreich, I.M. Gene-Environment interactions in stress response contribute additively to a genotype-environment interaction. *PLoS Genetics* **12**, e1006158 (2016).

45. McKhann, G. *et al.* Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939–44 (1984).

46. Messaoudi, S. *et al.* Endothelial Gata5 transcription factor regulates blood pressure. *Nature Communications* **6**, 8835 (2015).

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

48. Neale, B.M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genetics* **7**, e1001322 (2011).

49. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics* **86**, 832–838 (2010).

50. Schork, N.J., Murray, S.S., Frazer, K.A. & Topol, E.J. Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development* **19**, 212–219 (2009).

51. Shi, M., Umbach, D.M. & Weinberg, C.R. Using parental phenotypes in case-parent studies. *Frontiers in Genetics* **6** (2015).

52. Sinnwell, J.P., Therneau, T.M. & Schaid, D.J. The kinship2 R package for pedigree data. *Human Heredity* **78**, 91–93 (2014).

53. So, H.C., Kwan, J.S.H., Cherny, S.S. & Sham, P.C. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *American Journal of Human Genetics* **88**, 548–565 (2011).

54. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).

55. Thornton, T. & McPeek, M.S. Case-control association testing with related individuals: A more powerful quasi-likelihood score test. *American Journal of Human Genetics* **81**, 321–337 (2007).

56. Wacholder, S. *et al.* The kin-cohort study for estimating penetrance. *American Journal of Epidemiology* **148**, 623–630 (1998).

57. Wang, Y., Chen, H., Peloso, G.M., DeStefano, A. & Dupuis, J. Exploiting family history in aggregation unit-based genetic association tests. *bioRxiv*, 2021.04.05.438533 (2021).

58. Wessel, J. *et al.* Rare non-coding variation identified by large scale whole genome sequencing reveals unexplained heritability of type 2 diabetes. *medRxiv*, 2020.11.13.20221812 (2020).

59. Wu, M.C. *et al.* Rare-Variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* **89**, 82–93 (2011).

60. Zhao, Z.C. *et al.* UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. *American Journal of Human Genetics* **106**, 3–12 (2020).

61. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335–1341 (2018).

62. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nature Genetics* **52**, 634–639 (2020).

CURRICULUM VITAE

[redacted]

[redacted]

[redacted]