2015

Sequence Kernel Association Test, gene-environment interaction test, and meta-analysis for family samples with repeated measurements or multiple traits

Gao, Wei
SEQUENCE KERNEL ASSOCIATION TEST, GENE-ENVIRONMENT INTERACTION TEST, AND META-ANALYSIS FOR FAMILY SAMPLES WITH REPEATED MEASUREMENTS OR MULTIPLE TRAITS

by

WEI GAO

B.S., Fudan University, 2008
M.S., The George Washington University, 2010

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
2015
Approved by

First Reader
Josée Dupuis, Ph.D.
Professor of Biostatistics

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

Third Reader
Qiong Yang, Ph.D.
Associate Professor of Biostatistics
Acknowledgments

First, and foremost, I would like to give my sincere thanks to my major advisor Dr. Josée Dupuis for her guidance and encouragement throughout my journey towards this degree. I greatly appreciate her support and care. I would never have made it this far without her mentoring.

I would like to extend my thanks to the whole dissertation committee: Dr. L. Adrienne Cupples, Dr. Qiong Yang, Dr. Michael LaValley, and Dr. Ching-Ti Liu for their time and invaluable advice. I am also grateful to all my teachers for their help through my graduate study.

My PhD life would not be so joyful without the support from my friends. Special thanks to Dr. Han Chen for the discussions we had about my research, and to Vadim Sherman for helping me with my presentations.

Finally, I want to share the joy with my parents, Dr. Yongyi Gao and Guo Tang, who unconditionally love, support and believe in me.
SEQUENCE KERNEL ASSOCIATION TEST, GENE-ENVIRONMENT INTERACTION TEST, AND META-ANALYSIS FOR FAMILY SAMPLES WITH REPEATED MEASUREMENTS OR MULTIPLE TRAITS

( Order No.  )

WEI GAO

Boston University Graduate School of Arts and Sciences, 2015

Major Professor: Josée Dupuis, Professor of Biostatistics

ABSTRACT

Genetic loci identified by single variant association tests account for only a small proportion of the heritability for most complex traits and diseases. Part of the unexplained heritability may be due to rare variants and their interactions with environmental factors. Different strategies have been taken to increase the power to detect genetic associations, such as increasing the sample size by including related participants and meta-analyzing multiple studies. Longitudinal data or repeated measurements are often available in prospective cohort studies. For complex diseases, multiple traits are usually collected to characterize affected individuals. Many of the existing statistical methods can only be applied to the scenarios when each participant has one measurement of a single trait. To take full advantage of the data and further improve power, multiple measurements per individual may be included in the analysis when available. In this dissertation we develop statistical methods for rare variant association testing and gene by environment
interaction analysis, and discuss gene-based meta-analysis for studies with different
designs. First, we propose the generalized Sequence Kernel Association Test (genSKAT)
to deal with rare variants, familial correlation, and repeated measurements or multiple
traits. This is an extension of the original SKAT and family-based SKAT that accounts
for correlation between multiple measurements within each individual. In the second part
of this dissertation, we discuss methods to test for the presence of gene-environment
interaction effects in the genSKAT framework. Finally, we evaluate genSKAT meta-
analysis methods to combine different types of studies: samples of unrelated individuals
with one observation per person or with multiple observations per person, and family
samples with one observation per person or with multiple observations per person.
Combining all these projects together, we contribute methodologies to detect rare variant
associations by taking advantage of additional information, improve the chance to detect
novel rare variant associations, and help in understanding the role that genetic factors
play in various diseases and traits.
# Table of Contents

**Chapter 1**  Introduction ........................................................................................................ 1  
1.1  Genetics Association Studies .................................................................................... 1  
1.1.1  Rare Genetic Variants Analysis ................................................................. 3  
1.1.2  Gene-Environment Interaction .................................................................. 6  
1.1.3  Meta-analysis ......................................................................................... 8  
1.2  Genetics Association Studies for Multiple Measurements .............................. 9  
1.3  Dissertation Outline ..................................................................................... 11  

**Chapter 2**  Sequence Kernel Association Test for Family Samples with Repeated  
Quantitative Measurements or Multiple Quantitative Traits ............................. 13  
2.1  Introduction .................................................................................................. 13  
2.2  Method ....................................................................................................... 15  
2.3  Simulation Studies .................................................................................... 21  
2.3.1  Type I Error .......................................................................................... 21  
2.3.1.1  Simulation Design ........................................................................ 21  
2.3.1.2  Simulation Results ..................................................................... 24  
2.3.2  Power .................................................................................................... 29  
2.3.2.1  Simulation Design ........................................................................ 29  
2.3.2.2  Simulation Results ..................................................................... 30  
2.4  Application to Framingham Heart Study Data ............................................. 32  
2.5  Discussion .................................................................................................. 35  

vii
Chapter 3 Gene-Environment Interaction Test for Family Samples with Repeated Quantitative Measurements or Multiple Quantitative Traits

3.1 Introduction ........................................................................................................... 38
3.2 Methods .................................................................................................................. 40
3.3 Simulation Studies ................................................................................................. 47
  3.3.1 Type I Error......................................................................................................... 47
    3.3.1.1 Simulation Design ....................................................................................... 47
  3.3.2 Power ................................................................................................................ 53
    3.3.2.1 Simulation Design ....................................................................................... 53
      3.3.2.2 Simulation Results ................................................................................... 55
3.4 Application to Framingham Heart Study Data ..................................................... 58
3.5 Discussion .............................................................................................................. 61

Chapter 4 Gene-based Meta-analysis for Cohorts with Different Designs ............... 63
4.1 Introduction ............................................................................................................ 63
4.2 Method ................................................................................................................... 66
  4.2.1 Study Level Statistics ...................................................................................... 67
  4.2.2 Meta-analysis Assuming Homogeneity ............................................................. 69
  4.2.3 Meta-analysis Assuming Heterogeneity ......................................................... 70
4.3 Simulation Studies ............................................................................................... 73
  4.3.1 Type I Error ....................................................................................................... 73
  4.3.1.1 Simulation Design ....................................................................................... 73
List of Tables

Table 1 Inflated type I error for famSKAT ignoring correlations between multiple measurements .......................................................... Error! Bookmark not defined.

Table 2 Type I errors for genSKAT and other approaches (α=0.01) ............................................................................. 26
Table 3 Type I errors for genSKAT and other approaches (α=0.001) ............................................................................. 27
Table 4 Type I errors for genSKAT and other approaches (α=0.0001) ............................................................................. 28
Table 5 Application of genSKAT to FEV1/FVC .............................................................................................................. 34
Table 6 Application of famSKAT and famBT to FEV1/FVC ............................................................................................. 34
Table 7 Type I errors for genSKAT and other approaches (α=0.01, positive/negative/no effects=80%/0%/20%, 100,000 replicates) ........................................................................................................... 51
Table 8 Type I errors for genSKAT and other approaches (α=0.001, positive/negative/no effects=80%/0%/20%, 100,000 replicates) ........................................................................................................... 51
Table 9 Type I errors for genSKAT and other approaches (α=0.01, positive/negative/no effects=40%/40%/20%, 100,000 replicates) ........................................................................................................... 52
Table 10 Type I errors for genSKAT and other approaches (α=0.001, positive/negative/no effects=40%/40%/20%, 100,000 replicates) ........................................................................................................... 52
Table 11 Application of genSKAT for interaction effects to gene by pack-years ......................................................... 60
Table 12 Application of famSKAT and famBT for interaction effects to gene by sex ............................................. 60
Table 13 Different study designs ................................................................................................................................. 74
Table 14 Study designs for type I error simulations .................................................................................................. 78
Table 15 Type I errors for different meta-analysis approaches (α=0.01) ................................................................. 78
Table 16 Type I errors for different meta-analysis approaches (α=0.001) ................................................................. 79
Table 17 Type I errors for different meta-analysis approaches (α=0.0001) ................................................................. 79
List of Figures

Figure 1 Power comparisons of genSKAT, famSKAT, and famBT .............................. 31
Figure 2 Power for gene by continuous variable interaction effect tests ....................... 56
Figure 3 Power for the gene by binary variable interaction effect tests ....................... 57
Figure 4 Power comparison for genSKAT meta-analysis and other approaches ........... 84
List of Abbreviations

BT  Burden Test
CHARGE-S  Cohorts for Heart and Aging Research in Genomic Epidemiology
          Targeted Sequencing Project
Chr  Chromosome
COPD  Chronic Obstructive Pulmonary Disease
famBT  Family-Based Burden Test
famSKAT  Family-Based Sequence Kernel Association Test
FEV1  Forced Expiratory Volume in 1 Second
FEV1/FVC  Forced Expiratory Volume in 1 Second to Forced Vital Capacity Ratio
genSKAT  Generalized Sequence Kernel Association Test
GWAS  Genome-Wide Association Studies
G-E  Gene-environment
Het-Meta-SKAT  SKAT Meta-Analysis Assuming Heterogeneous Genetic Effects across Studies
Het-Meta-SKAT-O  SKAT-O Meta-Analysis Assuming Heterogeneous Genetic Effects across Studies
Hom-Meta-SKAT  SKAT Meta-Analysis Assuming Homogeneous Genetic Effects across Studies
Hom-Meta-SKAT-O  SKAT-O Meta-Analysis Assuming Homogeneous Genetic Effects across Studies
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD</td>
<td>Linkage Disequilibrium</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor Allele Frequency</td>
</tr>
<tr>
<td>SOLAR</td>
<td>Sequential Oligogenic Linkage Analysis Routines</td>
</tr>
<tr>
<td>SKAT</td>
<td>Sequence Kernel Association Test</td>
</tr>
<tr>
<td>SKAT-O</td>
<td>Sequence Kernel Association Test – Optimal Test</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>unrSKAT</td>
<td>Unrelated-Sample Sequence Kernel Association Test</td>
</tr>
<tr>
<td>SNV</td>
<td>Single Nucleotide Variant</td>
</tr>
<tr>
<td>VT</td>
<td>Variable-Threshold</td>
</tr>
</tbody>
</table>
Chapter 1   Introduction

1.1 Genetics Association Studies

In genetics association studies, we are interested in testing the effect of a specific allele on a given trait. Such association may be due to the functional effect of the genetic variant on the traits, or the genetic variant is in linkage disequilibrium with a functional genetic variant. Depending on whether the trait is qualitative or quantitative, different association tests are used. For qualitative traits, Chi-square tests or Fishers exact tests are commonly applied for case-control studies, and logistic regression is usually applied when adjustment for covariates is needed. For quantitative traits, linear regression is usually performed.

With the development of technology and the decreasing cost of genotyping, genome-wide association studies (GWAS) have been widely used to detect the association between genetic variants and complex diseases or traits. GWAS evaluate association with single nucleotide polymorphisms (SNPs) across the genome, which are common genetic variants defined by minor allele frequency (MAF) no less than 1%. In GWAS, a single variant association test, like logistic regression or linear regression, is performed for each
common variant. The p-value threshold of significance is adjusted for multiple testing.

Since the first successful study in 2005 [27], GWAS have identified thousands of SNPs related to multiple complex conditions including cardiovascular diseases, diabetes, and pulmonary diseases [1] [15] [19] [38] [47].

However, the genetic loci identified by single variant association tests account for only a small proportion of the heritability for most complex traits and diseases. Several reasons may account for the low heritability. First, part of the unexplained heritability may be due to rare variants [14]. Some large sample size studies revealed that the majority of genetic variants are rare variants [49] [59]. Some rare variants may have medical or biological effects on human diseases [9] [16] [17] [26] [40] [52]. In recent years, rapid advances in sequencing technologies, such as next-generation sequencing (NGS), provide opportunities to assess the effect of rare variants. At the same time, many statistical models for rare variant association analysis have been proposed, including the burden tests [2] [20] [34] [41] [45] [46] [51] [63], variance component tests [8] [36] [48] [50] [62], and combined tests [11] [24] [33]. Second, a portion of the variation in a given trait or disease may also be due to environmental factors and their interactions with genetic variants [18] [21] [53]. Many statistical methods for detecting gene-environment (G-E)
interaction have been proposed [7] [25] [28] [29] [35] [42] [44] [60] [61]. Third, some previous studies had in the order of hundreds of individuals. The relatively small sample size limits the power to detect the association of rare variants or variants with small effect sizes with traits [43]. Meta-analyzing data from a number of studies are usually applied to overcome this issue [3] [23] [32] [58] [64].

1.1.1 Rare Genetic Variants Analysis

Single variant tests usually have lower power to detect association with rare variants compared to common variant tests [34]. Many statistical methods have been proposed to improve power to detect association with rare variants [2] [8] [11] [17] [22] [24] [33] [34] [36] [41] [45] [46] [48] [50] [51] [62] [63]. These approaches usually evaluate the association between a trait and multiple rare variants in a gene or a region, instead of testing only one variant at a time. Power will be increased if multiple variants in the group are associated with the trait.

A typical statistical model for rare variants association tests is

\[ \mu(y) = X\beta + G\gamma + \epsilon, \]
where $\mu()$ is a link function, $X$ is a design matrix including covariates, $G$ is a genotype matrix of rare variants, $\beta$ and $\gamma$ are vectors of regression coefficients for $X$ and $G$ respectively. The null hypothesis of interest is

$$H_0: \gamma = 0$$

Burden tests [2] [34] [41] [45] [46] [63] collapse rare variants into a single genetic score by a weighted sum statistic. The association between this genetic score and the trait is tested. Burden tests are powerful when all variants have the same direction of effect. However, burden tests lose power substantially when most of the rare variants are null or rare variants have effects in opposite directions [4] [8] [33] [48]. To improve the power in the presence of null variants or bi-direction of effects, data-adaptive burden tests have been proposed. For the aSum test [20], negative weights are assigned to single nucleotide variants (SNVs) when the effect sizes are estimated to be negative. The variable-threshold (VT) method [51] optimizes the minor allele frequency threshold to select rare variants. Adaptive burden tests are more robust and more powerful than the original burden tests in the presence of null, protective and harmful variants. But usually p-values cannot be estimated analytically and such tests may require permutations to evaluate significance. Therefore many adaptive tests are computationally intensive.
The score-based variance component test is another class of rare variants association tests \[8\] \[48\] \[50\] \[62\]. In these tests, a score test statistic for a group of rare variants is created, and the distribution of the test statistic is evaluated. Instead of combing rare variants like burden tests, variance component tests use aggregated score test statistics. Variance component tests are more powerful than burden tests in the presence of rare variants with different direction of effects \[4\] \[8\] \[33\] \[48\] \[50\] \[62\], and are computationally efficient because permutation procedures are not required to evaluate statistical significance \[10\]. However variance component tests are less powerful than burden tests when most variants have the same direction of effects \[4\] \[8\] \[33\] \[48\] \[50\] \[62\]. The Sequence Kernel Association Test (SKAT), proposed by Wu et al. in 2011 \[62\], is a commonly used variance component test. It has been shown to be a powerful method in various scenarios. In 2013, Chen et al. extended the original SKAT to family-based SKAT (famSKAT) \[8\] for quantitative traits to account for familial correlation among related individuals.

Because of the pros and cons mentioned above, burden tests and variance component tests suit different scenarios. Some methods have been proposed to combine burden and
variance component test statistics. For instance, Lee et al. have proposed SKAT-O [33], an adaptive procedure to create a linear combination of SKAT and burden test statistics. Jiang et al. developed MONSTER [24], an extension of SKAT-O to accommodate familial correlation. The combined approaches are robust because the true association pattern between genetic variants and a given trait is usually unknown.

1.1.2 Gene-Environment Interaction

Both genetics and environmental exposures have effects on many phenotypes. There may also be some gene-environment interaction effects. For instance, individuals with specific genotypes may have different responses to lifestyle, treatment, drug dosages, or exposures. More genetic association studies are taking the interaction between genes and environment into consideration in recent years. There are two main purposes of including G-E interaction terms in a statistical model. First, the power to detect genetic effects may be increased when the gene-environment interaction exists and is correctly included in the model. Second, detecting a significant interaction effect itself may help in understanding the disease mechanism.
Usually, the statistical model for gene-environment interaction tests is

$$\mu(y) = X\beta + G\gamma + E\eta + \epsilon,$$

where $\mu()$ is a link function, $X$ is a design matrix including covariates, $G$ is a genotype matrix of rare variants, $E$ is a matrix for environment factors that are also included in $X$, and $\beta$, $\gamma$, and $\eta$ are vectors of regression coefficients for covariates, genotype, and G-E interaction respectively.

The null hypothesis to detect a genetic effect in the presence of G-E interaction is a joint test of

$$H_0: \gamma = \eta = 0.$$

The null hypothesis to detect a G-E interaction is

$$H_0: \eta = 0.$$

There are several methods to test interactions between common variants and environment [28] [29] [42] [44] [60]. Manning et al. proposed jointly estimating genetic main effects and G-E interaction effects for meta-analysis [42]. Moreno-Macias et al. discussed G-E interaction tests for a single measurement per person, e.g. OLR and FBAT, and for
longitudinal measures, e.g. OLMM and two-step modeling approach [44]. These methods usually lack power if directly applied to test interactions between rare variants and environment. Some set-based association testing methods have been proposed to improve the power of detecting G-E interaction for rare variants [7] [25] [35] [61]. Among them, Chen et al. discussed an interaction test for rare variants under the SKAT framework [7].

1.1.3 Meta-analysis

The rare variants association studies require large sample sizes to be well powered, especially when the effect size is not large [31]. Because of the high cost of sequencing and privacy issues, it is usually difficult to recruit more participants within a study. However, there may be different studies with the same traits available, and power may be increased when combining studies. Meta-analysis is often used to analyze data across studies to increase power because it does not require access to the individual level data.

One meta-analysis approach developed in the context of rare variants is based on combining test statistics instead of regression coefficients, because the regression coefficients may have very large variances due to the low allele frequency [32].
approach needs the minor allele frequency and score statistics for each rare variant, and
the covariance matrix of the score statistics from each study. Test statistic based meta-
analysis approaches can accommodate results from burden tests, SKAT, and SKAT-O.

1.2 Genetics Association Studies for Multiple Measurements

Different strategies have been taken to increase the power to detect genetic associations,
such as increasing the sample size by including additional study participants and meta-
analyzing multiple studies. Longitudinal data or repeated measurements are often
available in prospective cohort studies. Oftentimes, association between genetic variants
and a trait is tested using the average of multiple measurements, or using an observation
at a single time point per individual, such as the baseline or the end of the study. Useful
information may be lost during this data process. To take full advantage of the data and
further improve power, multiple measurements per individual should be included in the
analysis when available.

Multiple measurements per individual may be collected to better understand their
biological meaning. On the one hand, analyzing multiple measurements may reflect the
nature of the traits or diseases. For example, pulmonary functions are usually related to smoking status and pack-years, which may change over time. In pharmaceutical industry, the drug response is usually recorded over time or over an increase in dosage. Multiple measurements at different exposure levels may help in detecting the genetic association.

On the other hand, it is common to collect multiple traits to characterize affected individuals for complex diseases. Potential misclassification may happen due to divergence or ambiguity in the definition of complex diseases. The risk of misclassification may be reduced by joint analyses of related traits. For example, there are different ways to define a chronic obstructive pulmonary disease (COPD) case using a person’s forced expiratory volume in 1 second (FEV1) and FEV1 to forced vital capacity ratio (FEV1/FVC). A person may be diagnosed with COPD under one definition (e.g. FEV1 less than 80% of its predicted value and FEV1/FVC less than its lower limit of normal) but not under another definition (e.g. FEV1 less than its lower limit of normal and FEV1/FVC less than its lower limit of normal). Misclassification may exist in a genetic association study when using a binary variable of COPD as the outcome. Instead, jointly using the continuous measurements of FEV1 and FEV1/FVC as the outcomes of interest may minimize the misclassification issue.
1.3 Dissertation Outline

This dissertation includes 3 projects addressing different aspects of rare variant association studies. I developed statistical models to take advantage of multiple measurements in association studies. At the same time, I focus on methods that can be applied to family samples, thereby preventing the reduction in sample size when related individuals are omitted from analyses. This work involves rare variant association analysis, related individuals, repeated measurements or multiple traits, gene-environment interaction analysis, and meta-analysis.

In Chapter 2, I propose a general framework, genSKAT, to deal with rare variants, familial correlation, and repeated measurements or multiple traits based on the Sequence Kernel Association Test [8] [62]. To increase the power to detect associations, we may analyze multiple measurements per individual when available. For complex diseases, multiple traits are often collected to characterize affected individuals, and a joint analysis of related traits may provide increased power to detect genetic associations.
In Chapter 3, I discuss methods to test gene-environment interaction in the genSKAT framework. A portion of the variation in a given trait or disease may also be due to environmental factors and their interactions with genetic variants. I expect this approach to help in the understanding of the environmental factors that modify the genetic effects by using multiple measurements per individual.

In Chapter 4, I evaluate meta-analysis strategies to combine different types of cohorts: samples of unrelated individuals with one observation per person or with multiple observations per person, and family samples with one observation per person or with multiple observations per person. I perform simulation studies to estimate the type I error of genSKAT meta-analyses, and evaluate power when sample size and effect size change in different types of cohorts. I compare the power of genSKAT meta-analysis assuming homogeneous genetic effects, genSKAT meta-analysis assuming heterogeneous genetic effects, genSKAT on pooled data combining all individuals, and famSKAT meta-analysis assuming heterogeneity under different scenarios.

In Chapter 5, I summarize the research findings and discuss future work.
Chapter 2  Sequence Kernel Association Test for Family Samples with Repeated
Quantitative Measurements or Multiple Quantitative Traits

2.1 Introduction

Genetic loci identified by single variant association tests account for only a small
proportion of the heritability for most complex traits and diseases. Part of the unexplained
heritability may be due to rare variants [14]. Usually, single variant tests have low power
when applied to rare variants. Many statistical models for rare variant association
analysis have been proposed [2] [8] [11] [17] [22] [24] [33] [34] [36] [41] [45] [46] [48]
[50] [51] [62] [63]. Among them, the sequence kernel association test (SKAT) has been
shown to be a powerful method in various scenarios [62]. It is a score-based variance
component test that allows rare variants with different directions of effects, and is
computationally efficient because permutations are not required to evaluate statistical
significance [10]. The family-based SKAT (famSKAT) is an extension of SKAT [8] to
control the type I error by including familial correlation in the model.

There are three main strategies to increase the power to detect associations of rare
variants: 1) jointly analyzing multiple rare variants; 2) increasing the sample by including
additional individuals such as family members; and 3) analyzing multiple measurements
per individual when available. Longitudinal data or repeated measurements are often available in prospective cohort studies. For complex diseases, it is common to collect multiple traits to characterize affected individuals, and a joint analysis of related traits may provide increased power to detect genetic associations.

Using these three strategies, we introduce three layers of correlations: the correlation between variants, the correlation among family members, and the correlation between multiple traits or repeated measurements within each individual. Burden tests and the original SKAT adjust for the first layer of correlation, and famSKAT accounts for the first two layers of correlations. We propose a general SKAT framework (genSKAT) that takes all three layers of correlations into consideration in the test statistic formulation. Hence it is applicable to gene-based association studies with family samples, repeated phenotype measurements or multiple traits. When each participant has only one measurement, genSKAT is equivalent to SKAT when there is no familial correlation and to famSKAT in the presence of familial correlation.

SKAT and famSKAT will have inflated type I error if correlations between repeated phenotype measurements or multiple traits are inappropriately ignored. In our simulation
studies, we show that genSKAT has the correct type I error in the scenarios evaluated. GenSKAT has comparable power to famSKAT using mean phenotype data. We illustrate our approach to evaluate the association of rare genetic variants using longitudinal pulmonary function traits from the Framingham Heart Study data.

2.2 Method

We assume that there are $n$ individuals, some of them may be related to each other. Each individual has $m$ repeated measurements or $m$ traits. There are $p$ covariates and $q$ rare variants of interest. We assume that the quantitative traits follow a linear mixed effects model written in matrix form as

$$y = X\beta + G\gamma + \delta + \epsilon.$$ 

In this model, we assume that the correlation between multiple measurements is due to two sources, unmeasured genetic factors and environmental factors besides the covariates, represented by $\delta$ and $\epsilon$ respectively. More specific explanations of each term are below.
The vector \( \mathbf{y} = (y_1, y_2, \ldots, y_n)^T \) of size \( nm \times 1 \) contains the phenotype measurements, where \( y_i = (y_{i1}, y_{i2}, \ldots, y_{im}) \) and \( y_{ij} \) denotes the \( j^{th} \) measurement or trait of individual \( i \).

The matrix \( \mathbf{X} \) is an \( nm \times m(p + 1) \) design matrix of covariates.

\[
\mathbf{X} = \begin{pmatrix}
I_m & \mathbf{X}_1 \\
I_m & \mathbf{X}_2 \\
\vdots & \vdots \\
I_m & \mathbf{X}_n
\end{pmatrix},
\]

where \( I_m \) is an \( m \times m \) identity matrix, and

\[
\mathbf{X}_i = \begin{pmatrix}
x_{i1} & 0 & \ldots & 0 \\
0 & x_{i2} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & x_{im}
\end{pmatrix}, \mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp}), i = 1, 2, \ldots, n, j = 1, 2, \ldots, m.
\]

The vector \( \mathbf{x}_{ij} \) contains the measurements of \( p \) covariates for individual \( i \) at \( j^{th} \) phenotype measurement or trait. The matrix \( \mathbf{X} \) is written in this way to allow different intercepts and different covariate effects for different phenotype measurements or different traits.

The vector \( \boldsymbol{\beta} \) is an \( m(p + 1) \times 1 \) vector of fixed effect parameters,
\[ \beta = (\mu, \beta_1, \beta_2, \ldots, \beta_m)^T, \]

where \( \mu = (\mu_1, \mu_2, \ldots, \mu_m)^T \) contains \( m \) intercepts for \( m \) phenotype measurements or traits, and \( \beta_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})^T, j = 1,2, \ldots, m \) contains \( p \) fixed effects for \( j^{th} \) phenotype measurement or trait. There are a total of \( mp \) covariate effects. This number can be reasonably reduced to \( p \) covariate effects for repeated phenotype measurements, if we assume the fixed effect of the same covariate is the same for all repeated measurements.

The matrix \( G \) is an \( nm \times q \) genotype matrix for \( q \) rare variants. The genotype of each individual repeats \( m \) times (\( m \) rows) in this matrix. The \( y \) is a \( q \times 1 \) vector for the random effects of rare variants, \( \delta \) is an \( nm \times 1 \) vector for the unmeasured genetic effect, and \( \epsilon \) is an \( nm \times 1 \) vector for residuals. We assume \( y, \delta, \) and \( \epsilon \) are random, independent effects with

\[ y \sim N(0, \tau W), \]

\[ \delta \sim N(0, \Phi \otimes A), \]

\[ \epsilon \sim N(0, I_n \otimes R), \]
where $\tau$ is a corresponding variance component parameter, $W$ is a $q \times q$ diagonal weight matrix for the rare variants, $\Phi$ is an $n \times n$ relationship coefficient matrix or twice the kinship matrix, $I_n$ is an $n \times n$ identity matrix, and

$$A = \begin{pmatrix}
\sigma_{A_{11}}^2 & \sigma_{A_{12}} & \ldots & \sigma_{A_{1m}} \\
\sigma_{A_{12}} & \sigma_{A_{22}}^2 & \ldots & \sigma_{A_{2m}} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{A_{1m}} & \sigma_{A_{2m}} & \ldots & \sigma_{A_{mm}}^2
\end{pmatrix},
R = \begin{pmatrix}
\sigma_{E_{11}}^2 & \sigma_{E_{12}} & \ldots & \sigma_{E_{1m}} \\
\sigma_{E_{12}} & \sigma_{E_{22}}^2 & \ldots & \sigma_{E_{2m}} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{E_{1m}} & \sigma_{E_{2m}} & \ldots & \sigma_{E_{mm}}^2
\end{pmatrix}.$$  

Here $A$ is an $m \times m$ polygenic covariance matrix and $R$ is an $m \times m$ error covariance matrix for $m$ phenotype measurements or traits. The diagonal elements are the variances of each measurement due to unmeasured genetic information or error. The off-diagonal elements are the covariance between two measurements due to unmeasured genetic information or error. The symbol $\otimes$ denotes the Kronecker product. For multiple traits, some pairs of traits may be more correlated than others. The same applies for repeated measurements, where adjacent measurements are expected to be more correlated. Thus we use the Kronecker product to accommodate such correlation patterns.

We want to test whether a set of rare variants is associated with a quantitative phenotype.

In our parameter setting, we are interested in testing

$$H_0: \gamma = 0 \ vs. \ H_1: \gamma \neq 0.$$
As in Wu et al [62], we test the following null hypothesis instead:

\[ H_0: \tau = 0 \text{ vs. } H_1: \tau \neq 0. \]

The phenotypic variance is denoted as

\[ \Sigma = Var(y) = \tau G W G^T + \Phi \otimes A + I_n \otimes R. \]

Under the null hypothesis, \( y = X \beta + \delta + \epsilon \), estimates of the phenotypic variance and the effect of the genetic variants can be written as

\[ \hat{\Sigma} = \Phi \otimes \hat{A} + I_n \otimes \hat{R}, \]

\[ \hat{\beta} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} y. \]

The maximum likelihood estimators of \( \hat{A}, \hat{R} \), and \( \hat{\beta} \) are obtained under the null hypothesis. This is a major advantage of score tests because parameters do not need to be estimated under the alternative hypothesis, thus reducing the computational burden.

To obtain a variance component score test statistic, we need to take the derivative of the log likelihood for this linear mixed effects model with respect to \( \tau \). Following the same rationale as in the famSKAT statistic [8], the genSKAT score statistic can be written as
\[ Q = \left( y - X\hat{\beta} \right)^T \Sigma^{-1} GW G^T \Sigma^{-1} ( y - X\hat{\beta} ). \]

Under the null hypothesis,

\[ Q \sim \sum_{i=1}^{q} \lambda_i \chi^2_{1,i}, \]

where \( \lambda_i \) are the eigenvalues of \( W^{1/2} G^T \Sigma^{-1} P_0 \Sigma^{-1} G W^{1/2} \), with

\[ P_0 = Var( y - X\hat{\beta} ) = \Sigma - X ( X^T \Sigma^{-1} X )^{-1} X^T, \]

and \( \chi^2_{1,i} \) are independent \( \chi^2_1 \) random variables. The p-value can be computed by the Davies method [10] from the R package CompQuadForm [12] or the Kuonen’s saddlepoint method [30] from the R package survey [39]. The Davies’ method is slightly faster than the saddlepoint approximation. Chen et al. recommended the use of the Kuonen’s saddlepoint method when p-values are expected to be very small because the Davies’ method is less accurate for small p-values [8].

The null model and test statistic are in the same form as famSKAT, while the parameter settings are defined in a different way. When each individual only has 1 phenotype observation (\( m = 1 \)), genSKAT is equivalent to famSKAT. Additionally when there is no correlation between individuals, \( A = 0 \), genSKAT is equivalent to SKAT.
2.3   Simulation Studies

2.3.1  Type I Error

2.3.1.1 Simulation Design

We performed simulations to evaluate the type I error of genSKAT. We also included other tests, SKAT, famSKAT and the family based burden test (famBT), in our type-I error evaluation. For SKAT, we only included the unrelated individuals. For SKAT, famSKAT and famBT, we used baseline phenotype measurement, or the average of phenotype measurements. We also checked the type I error of famSKAT on full data without adjusting for the correlation between multiple measurements.

We simulated 100 genotype datasets for 250 nuclear families with 2 parents and 2 offspring, for a total of 1000 individuals per study. We included 20 variants with minor allele frequencies from 0.002 to 0.04. We first generated latent variables from the multivariate normal distribution with a first order autoregressive covariance structure. We set the parameter in the covariance structure to be 0.98, 0.9, and 0.8 to simulate the high, moderate, and low LD situations. Then the latent variables were dichotomized at specific quantiles decided by the minor allele frequencies to generate the haplotypes for founders.
The haplotypes were passed down to offspring without recombination according to Mendel’s segregation law.

Three scenarios were considered for the phenotype simulation: each individual had 2, 3, or 4 repeated phenotype measurements. For each scenario, we simulated 10,000 phenotype datasets with the formula

\[ y = 0.5x_1 + 0.5x_2 + \delta + \epsilon, \]

where \( x_1 \sim N(0, 1) \) and \( x_2 \sim Bernoulli(0.5) \) were two covariates. In real data, \( x_1 \) can be considered as a standardized continuous variable and \( x_2 \) can be considered as a binary variable like gender. The vectors \( \delta \) and \( \epsilon \) were generated from multivariate normal distributions \( N(0, \Phi \otimes A) \) and \( N(0, I_n \otimes R) \) respectively, where \( \Phi \) was the relationship coefficient matrix for the 250 nuclear families, and

\[
A = \begin{pmatrix} 1.60 & 0.96 \\ 0.96 & 1.60 \end{pmatrix}, \begin{pmatrix} 1.60 & 0.96 & 0.96 \\ 0.96 & 1.60 & 0.96 \\ 0.96 & 0.96 & 1.60 \end{pmatrix}, \begin{pmatrix} 1.60 & 0.96 & 0.96 & 0.96 \\ 0.96 & 1.60 & 0.96 & 0.96 \\ 0.96 & 0.96 & 1.60 & 0.96 \\ 0.96 & 0.96 & 0.96 & 1.60 \end{pmatrix},
\]

\[
R = \begin{pmatrix} 2.40 & 1.20 \\ 1.20 & 2.40 \end{pmatrix}, \begin{pmatrix} 2.40 & 1.20 & 0.96 \\ 1.20 & 2.40 & 1.20 \\ 0.96 & 1.20 & 2.40 \end{pmatrix}, \begin{pmatrix} 2.40 & 1.20 & 0.96 & 0.72 \\ 1.20 & 2.40 & 1.20 & 0.96 \\ 0.96 & 1.20 & 2.40 & 1.20 \\ 0.72 & 0.96 & 1.20 & 2.40 \end{pmatrix}.
\]
for each phenotype simulation scenario. Here we set the heritability to be 0.4, the phenotype variance to be 4, the pairwise genetic correlation to be 0.6, and the error correlation between adjacent measurements to be 0.5. We let the error correlation between farther measurements decrease by 0.1. The correlation matrix for 4 measurements was

\[
\begin{pmatrix}
1.00 & 0.54 & 0.48 & 0.42 \\
0.54 & 1.00 & 0.54 & 0.48 \\
0.48 & 0.54 & 1.00 & 0.54 \\
0.42 & 0.48 & 0.54 & 1.00
\end{pmatrix}
\]

Because we assumed the outcomes were repeated measurements of the same trait in these simulations, we set the unmeasured genetic variance and residual variance to be the same across measurements, which were the diagonal elements in \( A \) and \( R \). We also set the off-diagonal elements in matrix \( A \) to be the same, assuming the unmeasured genetic covariance between measurement pairs stayed constant. We allowed smaller residual covariance for measurements further apart in matrix \( E \). These assumptions were made to reduce the number of estimated parameters in our simulations. They were not required by the genSKAT method itself. Using unstructured covariance introduce more parameters and may cause the estimates to be unstable. For the type I error simulation, there were no genetic effects in this model (\( \gamma = 0 \)).
When evaluating p-values, we estimated $A$ and $R$ from the output of the Sequential Oligogenic Linkage Analysis Routines (SOLAR) software, and used the Wu weights for the weight matrix $W$. The empirical type I error of genSKAT and the other four methods were calculated at $\alpha$ levels of 0.01, 0.001 and 0.0001. The type I error was equal to the proportion of p-values less than or equal to the corresponding $\alpha$ level in the 1 million simulated datasets (100 genotypes $\times$ 10,000 phenotypes).

2.3.1.2 Simulation Results

FamSKAT has inflated type I error when it was applied to the full data without adjusting for correlations between multiple measurements per person (Table 1). Table 2-Table 4 shows the empirical type I errors when genSKAT applied to full data, SKAT applied to baseline or mean phenotype data of unrelated subjects, and famSKAT and famBT applied to baseline or mean phenotype data of family data. GenSKAT considered all 1,000 individuals with all data points. The total number of observations was 2,000, 3,000, or 4,000 when the number of measurements per person was 2, 3, or 4. FamSKAT and famBT was performed on all 1,000 individuals using baseline data or the mean of the measurements. The total number of observations was 1,000. The original SKAT only included unrelated individuals (unrSKAT), where the sample size and the total number of observations were 500. The type I errors were calculated in three different LD scenarios.
and three different numbers of observations per individual. In our simulations, we observe that genSKAT has correct type I errors. This is also true for unrSKAT, famSKAT and famBT using mean or baseline phenotype data.

<table>
<thead>
<tr>
<th>α level</th>
<th># of measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>0.01</td>
<td>0.04077</td>
</tr>
<tr>
<td>0.001</td>
<td>0.00854</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.00017</td>
</tr>
</tbody>
</table>
Table 2 Type I errors for genSKAT and other approaches (α=0.01)

<table>
<thead>
<tr>
<th>LD</th>
<th>Model</th>
<th># of measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>genSKAT</td>
<td>0.0908</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>0.0095</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0097</td>
</tr>
<tr>
<td></td>
<td>unrSKAT</td>
<td>0.0095</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0096</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0100</td>
</tr>
<tr>
<td>Moderate</td>
<td>genSKAT</td>
<td>0.0099</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>unrSKAT</td>
<td>0.0098</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0100</td>
</tr>
<tr>
<td>High</td>
<td>genSKAT</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>0.0098</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0099</td>
</tr>
<tr>
<td></td>
<td>unrSKAT</td>
<td>0.0096</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0097</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0098</td>
</tr>
</tbody>
</table>

The type I error was equal to the proportion of p-values less than or equal to the corresponding α level in the 1 million simulated datasets.
Table 3 Type I errors for genSKAT and other approaches (α=0.001)

<table>
<thead>
<tr>
<th>LD</th>
<th>Model</th>
<th># of measurements</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>genSKAT All</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famSKAT Baseline</td>
<td></td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>unrSKAT Baseline</td>
<td></td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>famBT Baseline</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>0.0010</td>
<td>0.0011</td>
<td>0.0010</td>
</tr>
<tr>
<td>Moderate</td>
<td>genSKAT All</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famSKAT Baseline</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>unrSKAT Baseline</td>
<td></td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>famBT Baseline</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td>High</td>
<td>genSKAT All</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famSKAT Baseline</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>unrSKAT Baseline</td>
<td></td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>famBT Baseline</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0011</td>
</tr>
</tbody>
</table>

The type I error was equal to the proportion of p-values less than or equal to the corresponding α level in the 1 million simulated datasets.
Table 4 Type I errors for genSKAT and other approaches (α=0.0001)

<table>
<thead>
<tr>
<th>LD</th>
<th>Model</th>
<th># of measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>genSKAT</td>
<td>0.00010</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.00010</td>
</tr>
<tr>
<td></td>
<td>unrSKAT</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.00010</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.00010</td>
</tr>
<tr>
<td>Moderate</td>
<td>genSKAT</td>
<td>0.00010</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>unrSKAT</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>0.00011</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.00011</td>
</tr>
<tr>
<td>High</td>
<td>genSKAT</td>
<td>0.00010</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>unrSKAT</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.00010</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>0.00009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.00010</td>
</tr>
</tbody>
</table>

The type I error was equal to the proportion of p-values less than or equal to the corresponding α level in the 1 million simulated datasets.
2.3.2 Power

2.3.2.1 Simulation Design

We compared the power of genSKAT, famSKAT and famBT. Repeated measurements were simulated, so that we could use the average phenotype measurement as the outcome variable for famSKAT and famBT. We applied genSKAT to the full data.

We generated 100 genotype datasets in the same way as for the type I error simulation. When generating the phenotype, we also considered three scenarios: each individual had 2, 3, or 4 phenotype measurements. For each scenario, 1,000 phenotype datasets were simulated with the equation

\[ y = 0.5x_1 + 0.5x_2 + G\gamma + \delta + \epsilon, \]

where \( x_1, x_2, \delta, \) and \( \epsilon \) were generated in the same way as in the type I error simulation. The vector \( \gamma \) contains the random effects of causal rare variants. We randomly selected 80\% of the rare variants as causal variants, and randomly assigned negative effect size to 0\% or 50\% of the causal variants. The absolute value of effect size \( \gamma_l \) for \( l^{th} \) variant is defined similarly as in Chen’s famSKAT paper [8],
\[ |\gamma| = \frac{c/v'Dv}{\sqrt{2MAF(1-MAF)^v}} \]

where \( c \) is a constant. In this formula, we set \( c \) to 0.05 when all causal variants’ effect sizes are positive and 0.5 when half of the causal variants’ effect sizes are negative. The matrix \( D \) is a first order auto-regressive matrix for the rare variants and \( v \) is a direction vector consisting of \( \pm 1 \).

When evaluating the p-values, we estimated the covariance matrices \( A \) and \( R \) from the output of the SOLAR software, and used the Wu weights in the weight matrix \( W \). We set the polygenic covariance between measurements to be the same. The power was equal to the proportion of p-values less than or equal to 0.001 in the 100,000 datasets.

2.3.2.2 Simulation Results

Figure 1 shows power simulation results for different numbers of repeated measurements per individual. The left panel in Figure 1 corresponds to the scenario when all causal rare variants have positive effect sizes. In this case, famBT is more powerful than famSKAT and genSKAT. The right panel in Figure 1 shows the results when half of the causal rare variants have negative effect sizes. FamBT has lower power because it cannot handle
variants with opposite directions of effect sizes well. FamSKAT and genSKAT do not dramatically lose power when there are opposite direction of interaction effects. If we don’t have prior information about the direction of genetic effects, we recommend genSKAT or famSKAT over famBT. GenSKAT and famSKAT on the average of measurements have similar power. The power of genSKAT and famSKAT increases with the number of repeated phenotype measurements.

Figure 1 Power comparisons of genSKAT, famSKAT, and famBT
2.4 Application to Framingham Heart Study Data

To show that genSKAT can handle missing values and complicated family structure in real data, we illustrated our approach by applying genSKAT to evaluate the association between pulmonary function score FEV1/FVC and genetic variants in two genes, ADAM19 and HTR4. Common variants rs2277027 and rs1422795 in ADAM19, and rs11168048 and rs7735184 in HTR4 have been reported to be associated to lung function measurements in a previous genome-wide association study [37]. We compared the results from genSKAT, famSKAT and famBT.

For ADAM19 and HTR4, we used the genotype data from Framingham Cohorts for Heart and Aging Research in Genomic Epidemiology Targeted Sequencing Project (CHARGE-S). Genes previously shown to be related to some phenotypes were selected for deep re-sequencing as part of the CHARGE-S project. We only analyzed polymorphic rare variants with MAF less than 5%. For ADAM19, we additionally restricted to be within ±1kb of the reported common variants rs2277027 (region 1) and rs1422795 (region 2). For HTR4, we included rare variants within ±6kb of rs11168048 (region 3) and ±7kb of rs7735184 (region 4). We set different gene region lengths in order to include at least 3 rare variants in each region.
We used phenotypes and covariates data from Framingham Heart Study offspring cohort at exams 5, 6, and 7. We performed genSKAT on all data points, while famSKAT and famBT were applied to the baseline data. Similar to a previous longitudinal CHARGE-S study on FEV1/FVC [57], we adjusted the analyses for age, sex, height, smoking status, pack-years, and principal components.

Table 5 and Table 6 show the p-values of the association tests for each region. We compare the p-values to an adjusted significance threshold for multiple tests by Bonferroni correction, which is 0.05/4=0.0125. For the 4 regions in ADAM19 and HTR4, no statistically significant associations are detected. Table 5 and Table 6 also show that far fewer observations were used by famSKAT and famBT compared to genSKAT. Note that not every individual has measurements at all 3 exams, so the number of observations is less than 3 times the total sample size in genSKAT. This application shows that genSKAT can handle missing values. The individuals in this Framingham Heart Study data also have more complicated familial correlations than the nuclear families in our simulation studies.
Table 5 Application of genSKAT to FEV1/FVC

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>Region</th>
<th>N</th>
<th># of observations</th>
<th># of variants</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM19</td>
<td>5</td>
<td>Region 1</td>
<td>952</td>
<td>2314</td>
<td>10</td>
<td>0.3173</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Region 2</td>
<td>952</td>
<td>2314</td>
<td>9</td>
<td>0.3224</td>
</tr>
<tr>
<td>HTR4</td>
<td>5</td>
<td>Region 3</td>
<td>952</td>
<td>2314</td>
<td>13</td>
<td>0.0134</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Region 4</td>
<td>952</td>
<td>2314</td>
<td>3</td>
<td>0.8583</td>
</tr>
</tbody>
</table>

Region 1: rs2277027±1kb; region 2: rs1422795±1kb; region 3: rs11168048±6kb; region 4: rs7735184±7kb.

Table 6 Application of famSKAT and famBT to FEV1/FVC

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>Region</th>
<th># of observations</th>
<th># of variants</th>
<th>famSKAT p-value</th>
<th>famBT p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM19</td>
<td>5</td>
<td>Region 1</td>
<td>952</td>
<td>10</td>
<td>0.8698</td>
<td>0.6305</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Region 2</td>
<td>952</td>
<td>9</td>
<td>0.6488</td>
<td>0.7293</td>
</tr>
<tr>
<td>HTR4</td>
<td>5</td>
<td>Region 3</td>
<td>952</td>
<td>13</td>
<td>0.8168</td>
<td>0.4621</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Region 4</td>
<td>952</td>
<td>3</td>
<td>0.2757</td>
<td>0.4589</td>
</tr>
</tbody>
</table>

Region 1: rs2277027±1kb; region 2: rs1422795±1kb; region 3: rs11168048±6kb; region 4: rs7735184±7kb.
2.5 Discussion

In our research, we proposed a general SKAT framework that can be used to analyze family data, multiple phenotype measurements or multiple traits. GenSKAT is an extension of SKAT [62] and famSKAT [8] which additionally accounts for correlations between multiple measurements within each individual. GenSKAT is equivalent to famSKAT when only 1 observation per individual is available, and is equivalent to SKAT if there is also no correlation between individuals. We showed that genSKAT has correct type I error for family data with repeated quantitative measurements or multiple quantitative traits in the scenarios evaluated. To avoid inflated type I error, famSKAT and famBT need to be applied to the baseline data or the average of phenotype measurements. Additionally, samples analyzed by SKAT need to be restricted to unrelated individuals. FamBT has lower power when causal variants have opposite direction of effects. GenSKAT and famSKAT do not suffer from such drastic power loss when causal variants have opposite direction of effects. In real data, when no prior information about the true direction of the genetic effects is available, we recommend genSKAT or famSKAT over burden tests. GenSKAT and famSKAT on the average of measurements have similar power in our simulations. In real data analyses, it is usually not appropriate to take the average of multiple traits or the average of some covariates,
e.g. smoking status and pack-years. In these cases, it is hard to interpret the results of applying famSKAT on the average of measurements.

We applied genSKAT to 4 gene regions that contain common variants previously shown to be associated with FEV1/FVC in large GWAS [37]. Because the common variants only explain part of the heritability, we investigated whether the rare variants in these genes are related to the trait. We used genotype data from the CHARGE-S project with rare variants from deep-sequenced targeted gene regions. We noted that the sample size in the Framingham Heart Study alone, even with repeated measurements, was much smaller than the sample size in the previous GWAS. We did not detect any significant associations between FEV1/FVC and the rare variants in these gene regions. In the Framingham Heart Study data, some phenotype or covariates measurements were missing for some individuals. The sample also had complex family structures. This application showed that genSKAT is a general and flexible approach that can be applied to complex data in real life.

Genetic sequencing has become faster and cheaper with the development of new technologies. Deep sequencing data are becoming available for cohorts with family
structure as well as longitudinal measurements. For instance, in Framingham Heart Study, there are participants from 3 generations. For each generation, there are multiple exams per person. With rare variants being sequenced from projects like CHARGE-S and Exome Chip projects, detection of novel gene regions associated with various traits using genSKAT becomes possible. GenSKAT may help to better understand diseases that are characterized by more than one phenotype. For example, as a complex disease, chronic obstructive pulmonary diseases (COPD) is usually diagnosed by FEV1 and FEV1/FVC. Joint analysis of these two traits in the same model using genSKAT may identify novel genes associated with COPD.

Similar to SKAT and famSKAT, genSKAT can be used to analyze common variants or a combination of common variants and rare variants. Using external weights for rare variants based on prior information can potentially further increase the power. The genSKAT statistic can be combined with other test statistics to create robust approaches like the SKAT-O method [33].
Chapter 3  Gene-Environment Interaction Test for Family Samples with Repeated Quantitative Measurements or Multiple Quantitative Traits

3.1 Introduction

As previously mentioned, the genetic loci identified by single variant association tests account for only a small proportion of the heritability for most complex traits and diseases. In Chapter 2, we discussed that part of the heritability could be due to rare variants and proposed a general framework for a rare variant association test called genSKAT. In this chapter, we discuss another possible explanation for the unexplained heritability, gene by environment interaction (G-E interaction). There have been studies showing G-E interaction effects contribute to a portion of phenotypic variation [18] [21] [53]. Different lifestyles, treatments, drug dosages, or exposures may have very different effects on the outcome of interest for individuals with different genotypes. For example, the protective effect of ADH7 variants for oesophageal cancer is due to an interaction with alcohol consumption [5]. There is an increasing interest in studying the interactive effects of environment factors and different genotypes.

Several statistical methods to test interactions between common variants and environment have been proposed. Manning et al. proposed jointly estimating genetic main effects and
G-E interaction effects for meta-analysis [42]. Moreno-Macias et al. discussed G-E interaction tests for a single measurement per person, e.g. OLR and FBAT, and for longitudinal measures, e.g. OLMM and two-step modeling approach [44]. The single variant G-E tests may lack power if directly applied to test interactions between a rare variant and environment. Chen et al. discussed G-E interaction tests and joint test of G and G-E interaction for rare variants under the SKAT framework [7]. In some studies, each individual may have more than one measurement, which provides more information. If we can take advantage of all available data points, we may be able to further improve the power of interaction tests. For complex diseases, the environment exposure and its interaction with genes may affect more than one trait. Moreover, the gene effects may be different at different time points. Therefore, a gene-environment interaction test under the genSKAT framework may offer improved power to identify genes playing an important role in trait regulation or disease risk.

There are two main purposes of including G-E interaction term in a statistical model. First, the power to detect a genetic effect may be increased when a gene-environment interaction exists and is correctly included in the model. In this situation, we jointly test the main effect of rare variants and the gene-environment interaction effect. Thus it is a
multi-dimensional test. Second, detecting a significant interaction effect itself may help in understanding the disease or trait mechanism. In this scenario, we are primarily interested in testing the interaction term.

In our simulation studies we showed that SKAT and famSKAT for G-E interaction effects obtain the correct type I error if being applied to the average of multiple available measurements or when using one observation per individual. The G-E interaction test in the genSKAT framework has the correct type I error in the scenarios evaluated. GenSKAT for G-E interaction effects has comparable power to famSKAT for G-E interaction effects using mean phenotype data. We illustrate our approach to evaluate G-E effects of rare genetic variants on pulmonary function traits from the Framingham Heart Study.

3.2 Methods

We assume that there are $n$ individuals, some of whom may be related, with $m$ repeated measurements or $m$ traits. There are $p$ covariates and $q$ rare variants of interest. Assume
the quantitative traits follow a linear mixed effects model that can be written in matrix form as

$$y = X\beta + G\gamma + EG\eta + \delta + \epsilon.$$

Correlation between multiple measurements is due to two sources, the unmeasured genetic factors and the environmental factors, which are represented by $\delta$ and $\epsilon$ respectively. In this model, following the same definitions as in section 2.2, $y$ is a $nm \times 1$ phenotype vector, $X$ is an $nm \times m(p + 1)$ covariate matrix, $\beta$ is an $m(p + 1) \times 1$ vector of fixed effect parameters, $G$ is an $nm \times q$ genotype matrix, $W$ is a $q \times q$ diagonal weight matrix for the rare variants, $\delta$ is an $nm \times 1$ vector for the unmeasured genetic effect, $\delta \sim N(0, \Phi \otimes A)$, and $\epsilon$ is an $nm \times 1$ vector for the residual, $\epsilon \sim N(0, I_n \otimes R)$.

Additionally, $E$ is a matrix for an environment factor that may have an interaction effect with the genotypes. We assume that the elements in matrix $E$ are also in the covariate matrix $X$. More specifically, $E$ is an $nm \times nm$ diagonal matrix,

$$E = diag(z_1, z_2, ..., z_n),$$

where $z_i = diag(x_{i1\nu}, x_{i2\nu}, ..., x_{i\nu m\nu})$, and $x_{ij\nu}$ is the $j^{th}$ measurement of $\nu^{th}$ covariate of individual $i$. We define $E$ in this way to keep our model in matrix format and
consistent with the model used to test main effects in Chapter 2. The matrix $E$ is a sparse matrix, so matrix operation functions designed for sparse matrices can be used in computing to increase efficiency and lower the memory usage.

The vector $\gamma$ of size $q \times 1$ contains the main effects of rare variants, which can be treated as fixed effects or random effects. If the main effects are treated as random effects, we assume $\gamma \sim N(0, \tau W)$, where $W$ is a $q \times q$ diagonal weight matrix for the rare variants and $\tau$ is the corresponding variance component parameter.

The vector $\eta$ of size $q \times 1$ contains the gene-environment interaction effects. We assume $\gamma, \eta, \delta$ and $\epsilon$ are independent, random effects with

$$\eta \sim N(0, \xi \tilde{W}),$$

where $\xi$ is the corresponding variance component parameter and $\tilde{W}$ is a $q \times q$ diagonal weight matrix for the rare variants. The matrix $\tilde{W}$ can be different from $W$. Usually it is reasonable to use the same weight matrix $W$ for both $\eta$ and $\gamma$ because the same set of rare variants are analyzed for both main effects and interaction effects.
In this project, we discuss the tests for the null hypothesis with respect to the interaction effect only. The hypotheses can be written as

\[ H_0: \eta = 0 \ vs. \ H_1: \eta \neq 0. \]

It is equivalent to the testing of

\[ H_0: \xi = 0 \ vs. \ H_1: \xi \neq 0. \]

Scenario 1. When the genetic main effects \( \gamma \) are treated as fixed effects, under the null hypothesis,

\[ y = X\beta + G\sqrt{W}\gamma + \delta + \epsilon. \]

We define two new matrices as

\[ H = (X, G\sqrt{W})_{nm \times [m(p+1)+q]}, \theta = (\beta, \gamma)_{[m(p+1)+q] \times 1}. \]

Then the model under null hypothesis in this scenario can be re-written as

\[ y = H\theta + \delta + \epsilon, \]
\( \delta \sim N(0, \Phi \otimes A), \)

\( \epsilon \sim N(0, I_n \otimes R). \)

The phenotypic variance is written as

\[ \Sigma = \text{Var}(y) = \xi (EG)W(EG)^T + \Phi \otimes A + I_n \otimes R. \]

Under the null hypothesis,

\[ \hat{\Sigma} = \Phi \otimes \hat{A} + I_n \otimes \hat{R}, \]

\[ \hat{\theta} = (H^T \hat{\Sigma}^{-1} H)^{-1} H^T \hat{\Sigma}^{-1} y. \]

The maximum likelihood estimates of \( \hat{A}, \hat{R}, \) and \( \hat{\theta} \) are obtained under the null hypothesis of no gene-environment interaction.

Following the same rationale as in the genSKAT main effect test statistic, the variance component score test statistic for G-E interaction effects can be written as

\[ Q_{\text{FIX}} = (y - H\hat{\theta})^T \hat{\Sigma}^{-1} (EG)W(EG)^T \hat{\Sigma}^{-1} (y - H\hat{\theta}). \]

Under the null hypothesis,
\[ Q_{FIX} \sim \sum_{i=1}^{q} \lambda_i \chi_{1,i}^2, \]

where \( \chi_{1,i}^2 \) are independent \( \chi_1^2 \) random variables and \( \lambda_i \) are the eigenvalues of the matrix

\[
W^{\frac{1}{2}} (EG)^T \hat{\Sigma}^{-1} P_0 \hat{\Sigma}^{-1} (EG) W^{\frac{1}{2}},
\]

with \( P_0 = Var (y - H\hat{\theta}) = \hat{\Sigma} - H (H^T \hat{\Sigma}^{-1} H)^{-1} H^T. \)

Scenario 2. If the genetic main effects \( y \) are treated as random effects, under the null hypothesis,

\[
y = X\beta + G\gamma + \delta + \epsilon,
\]

\[
\gamma \sim N(0, \tau W)
\]

\[
\delta \sim N(0, \Phi \otimes A),
\]

\[
\epsilon \sim N(0, I_n \otimes R).
\]

The phenotypic variance is

\[
\Sigma = Var(y) = \tau GWG^T + \xi (EG)W(EG)^T + \Phi \otimes A + I_n \otimes R.
\]

Under the null hypothesis,
\[ \hat{\Sigma} = \hat{\tau}GWG^T + \Phi \otimes \hat{A} + I_n \otimes \hat{R}. \]

\[ \hat{\beta} = \left( X^T \hat{\Sigma}^{-1} X \right)^{-1} X^T \hat{\Sigma}^{-1} y. \]

The maximum likelihood estimates of \( \hat{\tau}, \hat{A}, \hat{R}, \) and \( \hat{\beta} \) are obtained under the null hypothesis.

Following the same rationale as in the genSKAT main effect test statistic, the variance component score test statistic for a test of interaction can be written as

\[ Q_{RAN} = (y - X\hat{\beta})^T \hat{\Sigma}^{-1} (EG)W(EG)^T \hat{\Sigma}^{-1} (y - X\hat{\beta}). \]

Under the null hypothesis,

\[ Q_{RAN} \sim \sum_{l=1}^{q} \lambda_l \chi^2_{1,l}, \]

where \( \chi^2_{1,l} \) are independent \( \chi^2_1 \) random variables and \( \lambda_l \) are the eigenvalues of the matrix

\[ W^{\frac{1}{2}}(EG)^T \hat{\Sigma}^{-1} P_0 \hat{\Sigma}^{-1} (EG) W^{\frac{1}{2}}, \]

where \( P_0 = Var(y - X\hat{\beta}) = \hat{\Sigma} - X(X^T \hat{\Sigma}^{-1} X)^{-1} X^T. \)
The p-value can be computed by the Davies method [10] from the R package CompQuadForm [12] or the Kuonen’s saddlepoint method [30] from the R package survey [39]. Chen et al. recommended the latter because Davies’ method is less accurate for small p-values [8].

When each individual only has 1 observation, that is \( m = 1 \), the dimension of \( E \) is reduced to \( n \times n \). The test statistic and its distribution are the same as those in famSKAT for G-E interaction effects. Additionally when all individuals are unrelated, that is \( \delta = 0 \) or \( A = 0 \), the test statistic and its distribution are the same as those in the interaction test in SKAT.

3.3 Simulation Studies

3.3.1 Type I Error

3.3.1.1 Simulation Design

We performed simulations to evaluate the type I error of genSKAT for interaction effects and compare it with other methods. When we assume the genetic main effects are random effects, the test statistic and its distribution are in the same form as the scenario assuming
fixed genetic main effects. In our simulations we treated the genetic main effects as fixed
effects and assumed the outcomes were repeated measurements of the same trait. We
compared the type I error of genSKAT to famSKAT and famBT for interaction effects.
For genSKAT, we included all individuals with all data points. For famSKAT and
famBT, we used baseline phenotype measurement, or the average of repeated phenotype
measurements.

In the same way as in section 2.3.1.1, we simulated 100 genotype datasets for 500 nuclear
families with 2 parents and 2 offspring, for a total of 2,000 individuals per study. Three
scenarios were considered for the phenotype simulation: each individual had 2, 3, or 4
repeated phenotype measurements. For each scenario, we simulated 1,000 phenotype
datasets with the formula

$$y = 0.5x_1 + 0.5x_2 + Gy + \delta + \epsilon,$$

where $x_1 \sim N(0, 1)$ and $x_2 \sim Bernoulli(0.5)$ were two covariates. In real data, $x_1$ can be
considered as a standardized continuous variable, and $x_2$ can be considered as a binary
variable like gender. We assume there are genetic main effects but no interaction effects.
For the genetic main effects, we randomly selected 80% of the rare variants as causal
variants, and randomly assigned negative effect size to 0% or 50% of the causal variants. The effect sizes were generated in the same way as in section 2.3.2.1.

Using the same definition of $A$ and $R$ in section 2.3.1.1, $\delta$ and $\epsilon$ were generated from multivariate normal distribution $N(0, \Phi \otimes A)$, and $N(0, I_n \otimes R)$, where $\Phi$ is the relationship coefficient matrix or twice the kinship matrix of the 500 nuclear families. As mentioned previously, we assumed outcomes were repeated measurements of the same trait in these simulations. Therefore we set the unmeasured genetic variance and residual variance to be the same across measurements, which were the diagonal elements in $A$ and $R$. We also let the unmeasured genetic covariance between measurement pairs, the off-diagonal elements in $A$, stay constant. These assumptions were made to reduce the number of estimated parameters in our simulations. Theoretically they were not required by the genSKAT method itself. Using unstructured covariance introduce more parameters and may cause the estimates to be unstable. When computing the genSKAT statistic, we calculated $A$ and $R$ from the output of the Sequential Oligogenic Linkage Analysis Routines (SOLAR) software, and used the Wu weights for the both main effects and interaction effects. The empirical type I error of genSKAT and the other four methods were calculated at $\alpha$ levels of 0.01 and 0.001. The type I error was equal to the proportion
of p-values less than or equal to the corresponding $\alpha$ level in the 100,000 simulated datasets (100 genotype datasets $\times$ 1,000 phenotype datasets).

### 3.3.1.2 Simulation Results

Table 7-Table 10 show the empirical type I errors of genSKAT, famSKAT, and famBT for G-E interaction effects when all genetic main effects are positive or half of the genetic main effects are negative. Because all three methods account for familial correlations, they were performed on all individuals. FamSKAT and famBT were applied to the baseline data, and mean data, while genSKAT was applied to the repeated measurements. The type I errors were calculated for three different LD scenarios and three different numbers of observations per individual. Our simulation studies show that the interaction test in genSKAT has correct type I errors, as well as the interaction test in famSKAT and famBT using mean or baseline phenotype data.
Table 7 Type I errors for genSKAT and other approaches (α=0.01, positive/negative/no effects=80%/0%/20%, 100,000 replicates)

<table>
<thead>
<tr>
<th>LD</th>
<th>Model</th>
<th># of measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>genSKAT</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td>famSKAT Baseline</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td>famBT Baseline</td>
<td>0.0099</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0100</td>
</tr>
<tr>
<td>Moderate</td>
<td>genSKAT</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>famSKAT Baseline</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td>famBT Baseline</td>
<td>0.0098</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0098</td>
</tr>
<tr>
<td>High</td>
<td>genSKAT</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>famSKAT Baseline</td>
<td>0.0103</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td>famBT Baseline</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0100</td>
</tr>
</tbody>
</table>

Table 8 Type I errors for genSKAT and other approaches (α=0.001, positive/negative/no effects=80%/0%/20%, 100,000 replicates)

<table>
<thead>
<tr>
<th>LD</th>
<th>Model</th>
<th># of measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>genSKAT</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famSKAT Baseline</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famBT Baseline</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0010</td>
</tr>
<tr>
<td>Moderate</td>
<td>genSKAT</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>famSKAT Baseline</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famBT Baseline</td>
<td>0.0011</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0010</td>
</tr>
<tr>
<td>High</td>
<td>genSKAT</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famSKAT Baseline</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famBT Baseline</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.0010</td>
</tr>
</tbody>
</table>
Table 9 Type I errors for genSKAT and other approaches ($\alpha=0.01$, positive/negative/no effects=40%/40%/20%, 100,000 replicates)

<table>
<thead>
<tr>
<th>LD</th>
<th>Model</th>
<th># of measurements</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>genSKAT</td>
<td>All</td>
<td>0.0103</td>
<td>0.0102</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>Baseline</td>
<td>0.0103</td>
<td>0.0103</td>
<td>0.0103</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0103</td>
<td>0.0102</td>
<td>0.0103</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>Baseline</td>
<td>0.0104</td>
<td>0.0104</td>
<td>0.0104</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0103</td>
<td>0.0103</td>
<td>0.0103</td>
</tr>
<tr>
<td>Moderate</td>
<td>genSKAT</td>
<td>All</td>
<td>0.0101</td>
<td>0.0100</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>Baseline</td>
<td>0.0102</td>
<td>0.0102</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0102</td>
<td>0.0102</td>
<td>0.0101</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>Baseline</td>
<td>0.0104</td>
<td>0.0104</td>
<td>0.0104</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0103</td>
<td>0.0105</td>
<td>0.0104</td>
</tr>
<tr>
<td>High</td>
<td>genSKAT</td>
<td>All</td>
<td>0.0100</td>
<td>0.0101</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>Baseline</td>
<td>0.0102</td>
<td>0.0102</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0101</td>
<td>0.0100</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>Baseline</td>
<td>0.0099</td>
<td>0.0099</td>
<td>0.0099</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0100</td>
<td>0.0103</td>
<td>0.0100</td>
</tr>
</tbody>
</table>

Table 10 Type I errors for genSKAT and other approaches ($\alpha=0.001$, positive/negative/no effects=40%/40%/20%, 100,000 replicates)

<table>
<thead>
<tr>
<th>LD</th>
<th>Model</th>
<th># of measurements</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>genSKAT</td>
<td>All</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>Baseline</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0010</td>
<td>0.0009</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>Baseline</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0010</td>
<td>0.0011</td>
<td>0.0011</td>
</tr>
<tr>
<td>Moderate</td>
<td>genSKAT</td>
<td>All</td>
<td>0.0010</td>
<td>0.0009</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>Baseline</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>Baseline</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td>High</td>
<td>genSKAT</td>
<td>All</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>famSKAT</td>
<td>Baseline</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td>famBT</td>
<td>Baseline</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0011</td>
</tr>
</tbody>
</table>
3.3.2 Power

3.3.2.1 Simulation Design

To evaluate the power of genSKAT for interaction effects and compare it with other models, we performed simulation studies. When assume the genetic main effects are random effects, the test statistic and its distribution are in the same form as the scenario assuming fixed genetic main effects. In our simulations we treated the genetic main effects as fixed effects and assumed the outcomes were repeated measurements of the same trait. We compared the power of genSKAT, famSKAT and famBT for interaction effects. For genSKAT, we included all individuals with all data points. For famSKAT and famBT, we used the average of phenotype measurements as the outcome.

When simulating genotype, we set the LD between adjacent variants to be moderate because rare variants are rarely in high LD, and generated 100 genotype datasets in the same way as for the type I error simulation in section 3.3.1.1. When simulating phenotype, we still considered three scenarios: each individual had 2, 3, or 4 phenotype measurements. For each scenario, 1,000 phenotype datasets were simulated under two designs.
(1) To evaluate the interaction between genotype and a continuous covariate, we generate the phenotype using

\[ y = 0.5x_1 + 0.5x_2 + G\gamma + x_1G\eta + \delta + \epsilon. \]

(2) To evaluate the interaction between genotype and a dichotomous covariate, we generate the phenotype using

\[ y = 0.5x_1 + 0.5x_2 + G\gamma + (x_2 - 0.5)G\eta + \delta + \epsilon. \]

In these two scenarios, \( x_1, x_2, \delta, \) and \( \epsilon \) were generated in the same way as in the type I error simulation. The vector \( \gamma \) contains the main effects of causal rare variants. For \( \gamma \), we randomly selected 80% of the rare variants as causal variants, and randomly assigned negative effect size to 0% or 50% of the causal variants. The absolute value of main effect size \( \gamma_l \) was defined in the same way as in type I error simulations. The vector \( \eta \) contains the interaction effects of causal rare variants. For \( \eta \), we randomly assigned negative interaction effect sizes to 0% or 50% of the causal variants. The absolute value of interaction effect size \( \eta_l \) for \( l^{th} \) variant is defined as

\[ |\eta_l| = \sqrt{\frac{c\nu'D\nu}{2MAF_l(1-MAF_l)\text{Var}(x_k)}}, \quad k = 1 \text{ or } 2. \]

We used the same definitions of \( c, \nu \) and \( D \) as in section 2.3.2.1. The power was equal to the proportion of p-values less than or equal to \( 10^{-10} \) in the 100,000 datasets (100 genotype datasets \( \times \) 1,000 phenotype datasets).
3.3.2.2 Simulation Results

Figure 2 shows the power simulation results of testing the interaction between genotype and a continuous covariate $x_1$, when the number of repeated measurement per individual increases from 2 to 3 to 4. When all interaction effects are in the same direction, famBT for interaction effects is more powerful than famSKAT and genSKAT. When half of the interaction effects are negative, famBT for interaction effects has almost no power. FamSKAT and genSKAT for interaction effects do not dramatically lose power when there are opposite direction of interaction effects. With the number of repeated phenotype measurements increasing from 2 to 3 to 4, all three approaches gain power. FamSKAT and genSKAT have similar power in our simulations. Figure 3 shows the power simulation results of testing the gene by binary covariate $x_2$ interaction. The result in Figure 3 is similar to Figure 2. These methods have similar performance when analyzing gene by continuous variable interaction and gene by binary variable interaction. From both figures, we can observe that power depends on the proportion of negative interaction effects instead of the proportion of negative main effects.
Figure 2 Power for gene by continuous variable interaction effect tests

The genetic main effects are all positive. The “+/–0” above each plot represents the proportion of positive/negative/null interaction effects.

The genetic main effects are half positive and half negative. The “+/–0” above each plot represents the proportion of positive/negative/null interaction effects.
Figure 3 Power for the gene by binary variable interaction effect tests

The genetic main effects are all positive. The “+/−/0” above each plot represents the proportion of positive/negative/null interaction effects.

The genetic main effects are half positive and half negative. The “+/−/0” above each plot represents the proportion of positive/negative/null interaction effects.
3.4 Application to Framingham Heart Study Data

To show that genSKAT for interaction effects can handle missing values and complex family structure in real data, we illustrated our approach by applying genSKAT to evaluate the associations between FEV1/FVC and two interaction effects: gene by pack-years (continuous) interaction and gene by gender (binary) interaction. These two covariates are known to be associated with the pulmonary functions. We compared the p-values of genSKAT, famSKAT and famBT.

For ADAM19 and HTR4, we used the genotype data from Framingham Cohorts for Heart and Aging Research in Genomic Epidemiology Targeted Sequencing Project (CHARGE-S). Genes previously shown to be related to selected phenotypes were chosen for deep re-sequencing as part of the CHARGE-S project. Common variants rs2277027 and rs1422795 in ADAM19, and rs11168048 and rs7735184 in HTR4 have been reported to be related to lung function measurements in a previous genome-wide association study [37]. We analyzed polymorphic rare variants with MAF less than 5% only. For ADAM19, we included rare variants within ±1kb of the reported common variants rs2277027 (region 1) and rs1422795 (region 2). For HTR4, we included rare variants
within ±6kb of rs11168048 (region 3) and ±7kb of rs7735184 (region 4). We set different gene region lengths in order to include at least 3 rare variants in each region.

We used phenotypes and covariates from Framingham Heart Study offspring cohort at exams 5 and 6, and the third generation cohort at exams 1 and 2. We performed the genSKAT interaction effects model on all data points. FamSKAT and famBT interaction effects were applied to the baseline data. Similar to a previous longitudinal CHARGE-S study on FEV1/FVC [57], we adjusted the analyses for age, sex, height, smoking status, pack-years, and principal components.

Table 11 and Table 12 show the p-values of the association tests for each region. We compare the p-values to an adjusted significance threshold for multiple tests by Bonferroni correction, which is 0.05/4=0.0125. None of the tests reach statistical significance in regions 1, 3, and 4. There may be no interactions, or we do not have enough power to detect interaction effects. GenSKAT for the gene by sex interaction at region 2 reaches the Bonferroni adjusted significance level. It is not captured by famSKAT or famBT. This association could be true or a false positive. Further investigation is needed. This application shows that genSKAT can handle missing values
in real data. The participants in this Framingham Heart Study data also have more complex familial correlations than the nuclear families in our simulation studies.

Table 11 Application of genSKAT for interaction effects to gene by pack-years

<table>
<thead>
<tr>
<th>Gene</th>
<th>Region</th>
<th># of persons</th>
<th># of variants</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>genSKAT</td>
</tr>
<tr>
<td>ADAM19</td>
<td>Region 1</td>
<td>1104</td>
<td>11</td>
<td>0.0481</td>
</tr>
<tr>
<td></td>
<td>Region 2</td>
<td>1104</td>
<td>9</td>
<td>0.0196</td>
</tr>
<tr>
<td>HTR4</td>
<td>Region 3</td>
<td>1104</td>
<td>13</td>
<td>0.0576</td>
</tr>
<tr>
<td></td>
<td>Region 4</td>
<td>1104</td>
<td>2</td>
<td>0.5857</td>
</tr>
</tbody>
</table>

Region 1: rs2277027±1kb; region 2: rs1422795±1kb; region 3: rs11168048±6kb; region 4: rs7735184±7kb. For genSKAT, \( N_{obs} = 1927 \).

Table 12 Application of famSKAT and famBT for interaction effects to gene by sex

<table>
<thead>
<tr>
<th>Gene</th>
<th>Region</th>
<th># of persons</th>
<th># of variants</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>genSKAT</td>
</tr>
<tr>
<td>ADAM19</td>
<td>Region 1</td>
<td>1104</td>
<td>11</td>
<td>0.1591</td>
</tr>
<tr>
<td></td>
<td>Region 2</td>
<td>1104</td>
<td>9</td>
<td>0.0118</td>
</tr>
<tr>
<td>HTR4</td>
<td>Region 3</td>
<td>1104</td>
<td>13</td>
<td>0.3964</td>
</tr>
<tr>
<td></td>
<td>Region 4</td>
<td>1104</td>
<td>2</td>
<td>0.8107</td>
</tr>
</tbody>
</table>

Region 1: rs2277027±1kb; region 2: rs1422795±1kb; region 3: rs11168048±6kb; region 4: rs7735184±7kb. For genSKAT, \( N_{obs} = 1927 \).
3.5 Discussion

In our research, we discussed a gene-environment interaction test in the genSKAT framework. GenSKAT for G-E interaction effects is equivalent to the G-E interaction test in famSKAT when only 1 observation per individual, and equivalent to the G-E interaction test in SKAT if there is also no correlation between individuals.

We showed that genSKAT for interaction effects has the correct type I error for family data with repeated quantitative measurements or multiple quantitative traits in the scenarios evaluated. GenSKAT and famSKAT for interaction effects have comparable power in our simulations. FamBT for interaction effects is more powerful when interaction effects in the same direction and less powerful when the interaction effects in opposite direction. In real data, when information about the true direction of the G-E interaction effects is not available, we recommend the use of famSKAT or genSKAT instead of burden tests. In addition, the power of genSKAT, famSKAT and famBT for interaction effects does not depend on the proportion of negative main effects.
We applied genSKAT to 4 gene regions that contain common variants previously shown to be associated with FEV1/FVC in large GWAS [37]. We evaluated the interactions between the rare variants in these genes and two covariates, pack-years and gender. These two covariates are known to be associated pulmonary traits. Pack-years measures lifetime smoking, which is continuous. Gender is a binary variable. We noted that the sample size in Framingham Heart Study alone, even with the repeated measurements, was relatively small to detect gene by environment interactions. We did not find any significant gene by pack-years interaction effects in all 4 regions, or any gene by sex interaction effects in region 1, 3, and 4. GenSKAT provides a borderline significant p-value for gene by sex interaction at region 2. It is not captured by famSKAT or famBT. This association could be true or a false positive. Further investigation is needed. In our sample, some individuals had missing phenotype or covariates measurements. Our sample also has a complex family structure. This application shows that genSKAT for interaction effects is a general and flexible approach that can be applied to complex real data.
Chapter 4  Gene-based Meta-analysis for Cohorts with Different Designs

4.1 Introduction

For rare variants, large sample sizes are needed to have adequate power to detect associations [31]. However, sequencing studies to assess the association with rare variants often only have hundreds of participants, and thus may not be well powered. Because of the high cost of sequencing, privacy issues, and the rareness of some traits or diseases, it is may be difficult to recruit more participants within a study. Multiple studies with the same traits available may be combined to increase power. Meta-analysis is often used to combine results from multiple studies to increase sample size and power. It has some advantages over joint analysis of pooled data from all studies. First, only cohort-specific summary statistics are required for meta-analysis and individual level data are not needed. This approach avoids some ethical concerns with sharing of human genetics data. Also it is much easier to share summary statistic files because of the smaller file sizes compared to individual-level datasets. Second, meta-analysis allows studies to use a different set of covariates. It is usually difficult to perform joint analysis if different studies use different covariates. Also studies may have different study designs. For example, we can meta-analyze family studies with studies of unrelated individuals.
Different meta-analysis approaches depend on which assumption is made regarding the effect estimates across studies. Under the assumption of homogeneity of effect sizes, effect sizes are expected to be the same in all participating studies and only differ from each other due to sampling variation. If studies come from separate populations or have diverse environmental factors such as dietary or lifestyle behaviors, then genetic effect sizes may differ across studies, causing heterogeneity of effect sizes. In meta-analysis for rare variants, heterogeneity between studies may also be introduced by the different sequencing technologies and quality control processes.

In the context of rare variants, one meta-analysis strategy is to combine test statistics instead of regression coefficients, because the regression coefficients may have large variances due to the low allele frequency of the rare variants [32]. Lee et al. proposed a general framework for rare variants meta-analysis [32] for commonly used rare variant analysis methods, such as burden tests, SKAT, famSKAT, and SKAT-O. This approach uses the minor allele frequency and score statistics for each rare variant, and the covariance matrix of the score statistics from each study. The rare variant meta-analysis approach propose by Lee et al. is computationally efficient, because p-values can be calculated analytically.
GenSKAT is an extension of SKAT and famSKAT. The genSKAT test statistic has similar structure and distribution as the SKAT and famSKAT statistics. Extension of the meta-analysis framework to accommodate genSKAT is feasible. In meta-analyses, different studies may have different study designs. Some studies may consist of unrelated individuals, while others may have family samples, and both types of studies may also include longitudinal data or repeated measurements. The genSKAT approach we proposed in chapter 2 is a flexible method that can be applied to all these types of study designs.

In this chapter, we discuss meta-analyses of genetic main effects. We evaluate genSKAT meta-analyses with simulations to examine type I error for different combinations of study designs. We note that there is no previous research evaluating type I error of famSKAT meta-analysis. So we also estimate the type I error in our simulations for famSKAT using baseline data when repeated measures are available.
We perform simulation studies to evaluate the power of genSKAT meta-analysis when sample sizes and effect sizes may differ between studies. We compare the power of genSKAT meta-analysis assuming homogeneous genetic effects, genSKAT meta-analysis assuming heterogeneous genetic effects, genSKAT joint analysis on pooled data combining all individuals, and famSKAT meta-analysis using baseline measurements assuming heterogeneity under different scenarios.

4.2 Method

We assume that there are $K$ studies to be meta-analyzed. For the $k^{th}$ study, where $k = 1, 2, \ldots, K$, we consider the same linear mixed effects model as for genSKAT,

$$ y^k = X^k \beta^k + G^k \gamma^k + \delta^k + \epsilon^k, $$

$$ \gamma^k \sim N(0, \tau^k W^k), $$

$$ \delta^k \sim N(0, \Phi^k \otimes A^k), $$

$$ \epsilon^k \sim N(0, I_n \otimes R^k), $$

where $y^k, X^k, \beta^k, G^k, \tau^k, W^k, \Phi^k, A^k$ and $R^k$ are defined in the same way as $y, X, \beta, G, \tau, W, \Phi, A$, and $R$ in section 2.2, but with a superscript $k$ indicating the study number.
Participating studies may have different sets of covariates, different numbers of repeated measurements, and various family structures. So parameters $y^k, X^k, \beta^k, G^k, W^k$, $\Phi^k, A^k$ and $R^k$ may have different dimensions in different studies.

### 4.2.1 Study Level Statistics

The genSKAT test statistic for testing the null hypothesis

$$H_0: y^k = 0 \ vs. \ H_1: y^k \neq 0,$$

or equivalently in our model settings,

$$H_0: \tau^k = 0 \ vs. \ H_1: \tau^k \neq 0.$$

can be written as

$$Q^k = (y^k - X^k \hat{\beta}^k)^T \Sigma^{-1} G^k W^k G^T \Sigma^{-1} (y^k - X^k \hat{\beta}^k).$$

The un-weighted score statistic for study $k$ is

$$S^k = G^T \Sigma^{-1} (y^k - X^k \hat{\beta}^k),$$

where $\Sigma^k$ and $\hat{\beta}^k$ are maximum likelihood estimators of $\Sigma^k$ and $\beta^k$, respectively.
\[ \bar{\Sigma}^k = \Phi^k \otimes \hat{A}^k + I_n \otimes \hat{R}^k, \]

\[ \hat{\beta}^k = (X^k \bar{\Sigma}^{-1} X^k)^{-1} X^k \bar{\Sigma}^{-1} y^k. \]

The variance of the un-weighted score statistic \( S^k \) is

\[
V^k = G^k \bar{\Sigma}^{-1} \text{Var}(y^k - X^k \hat{\beta}^k) \bar{\Sigma}^{-1} G^k \\
= G^k \bar{\Sigma}^{-1} (\bar{\Sigma} - X^k \text{Var}(\hat{\beta}^k) X^k) \bar{\Sigma}^{-1} G^k \\
= G^k \bar{\Sigma}^{-1} (\bar{\Sigma} - X^k (X^k \bar{\Sigma}^{-1} X^k)^{-1} X^k) \bar{\Sigma}^{-1} G^k \\
= G^k \left( \bar{\Sigma}^{-1} - \bar{\Sigma}^{-1} X^k (X^k \bar{\Sigma}^{-1} X^k)^{-1} X^k \bar{\Sigma}^{-1} \right) G^k.
\]

In each study, we record \( S^k, V^k \) and the minor allele frequency of each rare variable.

These quantities will be used for meta-analysis. We discuss two types of meta-analysis,

one approach when genetic main effects are assumed to be homogeneous across studies,

and another approach when genetic main effects are assumed to be heterogeneous across studies.
4.2.2 Meta-analysis Assuming Homogeneity

First we discuss meta-analysis tests assuming that the effect sizes of genetic variants are the same across studies. To perform a meta-analysis under the homogeneity assumption, we combine the score statistics as Lee et al. proposed [32]:

$$Q_{hom} = \sum_{i=1}^{q} \left( \sum_{k=1}^{K} w_i^k s_i^k \right)^2.$$

Because of the homogeneity assumption, we usually use the same weight across studies for each rare variant based on its average minor allele frequency or prior information. That is, $w_i^k = w_i$ for $k = 1, 2, ..., K$. For example, if we use the Wu weights for rare variants, $w_i = Beta(MAF_i, 1, 25)$, the $MAF_i$ will be the average minor allele frequency across all studies. When using the same weights for all studies, we can rewrite the $Q_{hom}$ in matrix form, as described below.

Within each gene region, we sum up the weighted score vectors and un-weighted covariance matrices across all studies

$$S_{hom} = \sqrt{W} \sum_{k=1}^{K} S^k, \quad V_{hom} = \sum_{k=1}^{K} V^k,$$
where $\mathbf{W} = diag(w_1^2, w_2^2, \ldots, w_q^2)$ is a diagonal weight matrix for $q$ rare variants. Then

$$Q_{\text{hom}} = S_{\text{hom}}^T S_{\text{hom}}.$$  

Using the same argument we used to estimate the distribution of genSKAT statistic, we find that under the null hypothesis,

$$Q_{\text{hom}} \sim \sum_{j=1}^{q} \lambda_{\text{hom},j} \chi_{1,j}^2,$$

where $\lambda_{\text{hom},j}$ are the eigenvalues of $\sqrt{\mathbf{WV}_{\text{hom}}} \sqrt{\mathbf{W}}$, and $\chi_{1,j}^2$ are independent $\chi_1^2$ random variables. The p-value can be computed using the Davies method [10] from the R package CompQuadForm [12] or the Kuonen’s saddlepoint method [30] from the R package survey [39]. The Davies’ method is slightly faster than the saddlepoint approximation. Chen et al. recommended the use of the Kuonen’s saddlepoint method when p-values are expected to be very small because the Davies’ method is less accurate for small p-values [8].

4.2.3 Meta-analysis Assuming Heterogeneity

Genetic main effects may be heterogeneous across studies. To derive a meta-analysis statistic that allows for heterogeneity, we assume that the effect sizes of rare variants in
different studies are independent and follow the same distribution. The meta-analysis statistic is given by [32]

\[ Q_{het} = \sum_{i=1}^{q} \sum_{k=1}^{K} w_i^k S_i^k, \]

where \( w_i^k \) is the study-specific weight for \( i^{th} \) variant in \( k^{th} \) study, and \( S_i^k \) is the study-level score for \( i^{th} \) variant in \( k^{th} \) study.

We can re-write \( Q_{het} \) in a matrix form as described below. The cumulative score vector for \( q \) variants can be given by

\[ S_{het} = \sqrt{W^*S^*}, \]

where \( W^* = diag(W^1, W^2, \ldots, W^K) \), and \( S^* = (S^1, S^2, \ldots, S^K)^T \). Then

\[ Q_{het} = S_{het}^T S_{het}. \]

Under the null hypothesis,

\[ Q_{het} \sim \sum_{j=1}^{q} \lambda_{het,j} \chi^2_{1,j}, \]
Where $\chi^2_{1,j}$ are independent $\chi^2$ random variables, and $\lambda_j$ are the eigenvalues of $\sqrt{W^*V^*W^*}$, with

$$V^* = Var(S^*) = diag(V^1, V^2, \ldots, V^K).$$


Sometime, studies may belong to subgroups according to prior information. For example, a meta-analysis project may group studies by their geographic locations and ancestries. In each subgroup, the assumption of homogeneity may hold. In this case, we can first apply a meta-analysis test assuming homogeneity within each subgroup and then perform meta-analysis assuming heterogeneity across subgroups.
4.3 Simulation Studies

4.3.1 Type I Error

4.3.1.1 Simulation Design

We performed simulations to estimate the type I error of genSKAT meta-analyses statistic assuming either homogeneity or heterogeneity in genetic main effects. We noticed that there was no previous research evaluating the type I error of famSKAT meta-analysis. Therefore, we also estimated the type I error for famSKAT meta-analysis using the baseline measurements in our simulations.

We considered 4 different study designs: A, B, C and D (Table 13). The samples of study designs A and C consisted of unrelated individuals, and the samples of study designs B and D consisted of nuclear families with two parents and two offspring. Additionally, each individual in study designs A and B had only 1 measurement, while each individual in study designs C and D had 3 measurements. GenSKAT can be directly applied to these 4 study designs. FamSKAT can be directly performed on study designs A and B, but only baseline or average outcome measurements can be used in the case of study designs C or D.
Table 13 Different study designs

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Number of Repeated Measurements</th>
<th>Family size</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

We considered two different scenarios for the sample size in each study. The first scenario consists of all studies having the same number of individuals. In our simulations, we set the number of individuals to be 1,000 for each study. This means that the numbers of observations for study designs A and B are both 1,000, and the numbers of observations for study designs C and D are both 3,000. The second scenario consists of all studies having the same number of observations. In our simulations, we controlled the total number of observations to be 3,000 for all 4 study designs. This means 3,000 unrelated individuals for study design A, and 750 nuclear families with 2 parents and 2 offspring for study design B. For each scenario of the sample size, we evaluated the type I errors in three scenarios: meta-analyzing 4 type D studies, meta-analyzing 10 type D studies, and meta-analyzing 4 studies, one from each of type A, B, C and D respectively.
We generated 1,000 genotype datasets for each study. For studies with family samples, the genotype simulations were similar as in section 2.3.1.1. We assumed 18 variants with minor allele frequencies from 0.002 to 0.05. We first generated latent variables from the multivariate normal distribution with a first order autoregressive covariance structure. Then the latent variables were dichotomized at specific quantiles decided by the minor allele frequencies to generate the haplotypes for founders. The haplotypes were passed down to offspring without recombination according to Mendel’s segregation law. For studies with unrelated individuals, we followed the same genotype simulation process but stopped after generating haplotypes for founders.

For each genotype dataset, we generated 1,000 phenotype datasets. We simulated the continuous phenotype using the formula

\[ y^k = 0.5x_1^k + 0.5x_2^k + \delta^k + \epsilon^k, \]

where \( x_1^k \sim N(0, 1) \) and \( x_2^k \sim Bernoulli(0.5) \), and \( \delta^k \) and \( \epsilon^k \) are generated from multivariate normal distribution \( N(0, \Phi^k \otimes A^k) \) and \( N(0, I_n \otimes R^k) \) respectively. The matrix \( \Phi^k \) is the relationship coefficient matrix for study designs B and D, and the identity matrix for study designs A and C. \( A^k \) and \( R^k \) were defined by
These definitions followed the same rationale as in section 2.3.1.1. There are no genetic effects in this model because we are evaluating type I error simulation under null hypothesis.

When evaluating p-values, we used Wu weights for both genSKAT meta-analyses assuming homogeneity and assuming heterogeneity, although minor allele frequencies (MAFs) used vary depending on the homogeneity or heterogeneity assumption. For meta-analyses assuming homogeneity, we calculated the Wu weights on average MAFs from all studies, and used them as weights for all studies. For meta-analyses assuming heterogeneity, we calculated study-specific weights with the study-specific MAFs. The empirical type I error was calculated at $\alpha$ levels of 0.01, 0.001 and 0.0001. The type I error was equal to the proportion of p-values less than or equal to the corresponding $\alpha$ level in the 1 million datasets (1,000 genotype datasets $\times$ 1,000 phenotype datasets).
4.3.1.2 Simulation Results

The type I error simulation results for genSKAT meta-analysis assuming homogeneity or heterogeneity and other methods are displayed in Table 15 to Table 17. All repeated measurements were included in genSKAT meta-analyses. Only the baseline measure was used in famSKAT meta-analyses. We evaluated type I errors for 6 scenarios with different sample sizes and different study designs. Table 14 shows the number of observations and number of individuals in each study for the 6 scenarios. As seen in Table 15 to Table 17, the type I error of all methods is well controlled for all significance levels evaluated.
### Table 14 Study designs for type I error simulations

<table>
<thead>
<tr>
<th>Sample size scenario</th>
<th>Study design</th>
<th>$N_{\text{individuals}}$</th>
<th>$N_{\text{observations}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same number of</td>
<td>4 × D</td>
<td>1,000 for each study</td>
<td>3,000 for each study</td>
</tr>
<tr>
<td>individuals</td>
<td>10 × D</td>
<td>1,000 for each study</td>
<td>3,000 for each study</td>
</tr>
<tr>
<td></td>
<td>A, B, C, D</td>
<td>1,000 for each study</td>
<td>1,000, 1,000, 3,000, 3,000</td>
</tr>
<tr>
<td>Same number of</td>
<td>4 × D</td>
<td>1,000 for each study</td>
<td>3,000 for each study</td>
</tr>
<tr>
<td>observations</td>
<td>10 × D</td>
<td>1,000 for each study</td>
<td>3,000 for each study</td>
</tr>
<tr>
<td></td>
<td>A, B, C, D</td>
<td>3,000, 3,000, 1,000, 1,000</td>
<td>3,000 for each study</td>
</tr>
</tbody>
</table>

### Table 15 Type I errors for different meta-analysis approaches ($\alpha=0.01$)

<table>
<thead>
<tr>
<th>Sample size scenario</th>
<th>Study design</th>
<th>Method</th>
<th>Method</th>
<th>Method</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>genSKAT Hom Meta</td>
<td>genSKAT Het Meta</td>
<td>famSKAT Hom Meta</td>
<td>famSKAT Het Meta</td>
</tr>
<tr>
<td>Same number of</td>
<td>4 × D</td>
<td>0.0097</td>
<td>0.0098</td>
<td>0.0101</td>
<td>0.0102</td>
</tr>
<tr>
<td>individuals</td>
<td>10 × D</td>
<td>0.0097</td>
<td>0.0099</td>
<td>0.0101</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td>A, B, C, D</td>
<td>0.0099</td>
<td>0.0099</td>
<td>0.0102</td>
<td>0.0103</td>
</tr>
<tr>
<td>Same number of</td>
<td>4 × D</td>
<td>0.0096</td>
<td>0.0097</td>
<td>0.0101</td>
<td>0.0101</td>
</tr>
<tr>
<td>observations</td>
<td>10 × D</td>
<td>0.0098</td>
<td>0.0010</td>
<td>0.0100</td>
<td>0.0099</td>
</tr>
<tr>
<td></td>
<td>A, B, C, D</td>
<td>0.0010</td>
<td>0.0099</td>
<td>0.0100</td>
<td>0.0101</td>
</tr>
</tbody>
</table>
Table 16 Type I errors for different meta-analysis approaches (α=0.001)

<table>
<thead>
<tr>
<th>Sample size scenario</th>
<th>Study design</th>
<th>genSKAT Hom Meta</th>
<th>genSKAT Het Meta</th>
<th>famSKAT Hom Meta</th>
<th>famSKAT Het Meta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same number of individuals</td>
<td>4 × D</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0011</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>10 × D</td>
<td>0.0010</td>
<td>0.0009</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>A, B, C, D</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0010</td>
<td>0.0011</td>
</tr>
<tr>
<td>Same number of observations</td>
<td>4 × D</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>10 × D</td>
<td>0.0009</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>A, B, C, D</td>
<td>0.0010</td>
<td>0.0009</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

Table 17 Type I errors for different meta-analysis approaches (α=0.0001)

<table>
<thead>
<tr>
<th>Sample size scenario</th>
<th>Study design</th>
<th>genSKAT Hom Meta</th>
<th>genSKAT Het Meta</th>
<th>famSKAT Hom Meta</th>
<th>famSKAT Het Meta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same number of individuals</td>
<td>4 × D</td>
<td>0.00010</td>
<td>0.00009</td>
<td>0.00011</td>
<td>0.00010</td>
</tr>
<tr>
<td></td>
<td>10 × D</td>
<td>0.00010</td>
<td>0.00010</td>
<td>0.00010</td>
<td>0.00010</td>
</tr>
<tr>
<td></td>
<td>A, B, C, D</td>
<td>0.00009</td>
<td>0.00009</td>
<td>0.00010</td>
<td>0.00011</td>
</tr>
<tr>
<td>Same number of observations</td>
<td>4 × D</td>
<td>0.00009</td>
<td>0.00009</td>
<td>0.00011</td>
<td>0.00011</td>
</tr>
<tr>
<td></td>
<td>10 × D</td>
<td>0.00009</td>
<td>0.00010</td>
<td>0.00010</td>
<td>0.00010</td>
</tr>
<tr>
<td></td>
<td>A, B, C, D</td>
<td>0.00010</td>
<td>0.00010</td>
<td>0.00010</td>
<td>0.00010</td>
</tr>
</tbody>
</table>
4.3.2 Power

4.3.2.1 Simulation Design

We performed simulations to evaluate the power of genSKAT meta-analyses, assuming homogeneity or heterogeneity in genetic main effects. We compared the power to genSKAT joint analysis on pooled data, and to famSKAT meta-analysis assuming heterogeneity. Similar to the type I error simulations, we were interested in 4 different study designs, A, B, C and D defined in Table 13. We considered two different scenarios of the sample size: all studies with the same number of individuals, and all studies with the same number of observations. For each sample size scenario, we evaluated the power of meta-analyzing 4 studies, 1 from each study design A, B, C, and D.

We generated 1,000 genotype datasets in the same way described in the type I error simulation section. For each of 100,000 iterations, 4 genotype datasets were randomly sampled from the 1,000 genotype datasets and were assigned to be the genotypes for 4 studies of type A, B, C, and D respectively. For each set of 4 genotype datasets, we generated the continuous phenotypes using the formula

\[ y^k = G^k \gamma^k + \delta^k + \epsilon^k, \; k = 1,2,3,4, \]
where $\delta^k$ and $\varepsilon^k$ were generated from multivariate normal distribution $N(0, \Phi^k \otimes A^k)$ and $N(0, I_n \otimes R^k)$ respectively, and $G^k$ was the genotype matrix, and $\gamma^k$ contained the random effects of rare variants. $\Phi^k$ was the relationship coefficient matrix for study design B or D, and the identity matrix for study design A or C. $A^k$ and $R^k$ were defined in the same way as in type I error simulations. We did not include any covariates in the model. In meta-analyses on real data, each study may run study-level models adjusting for covariates and generate residuals. Then the residuals from each study can be treated as the new outcome.

We randomly selected 50% of the rare variants as causal variants, and randomly assigned negative effect size to half of the causal variants. The absolute value of effect size $\gamma_l$ was defined similarly as in section 2.3.2.1. We want to evaluate power of the meta-analyses methods in three effect size scenarios: 1) all studies have the same effect sizes; 2) studies with a single measurement per person (types A and B) have half of the effect sizes of studies with repeated measurements (types C and D); and 3) study types C and D have half of the effect sizes of study types A and B. When calculating the Wu weights for rare variants, we used the average MAFs for meta-analyses assuming homogeneity, and used the study-specific MAFs for meta-analyses assuming heterogeneity.
4.3.2.2 Simulation Results

The left panel from Figure 4 shows the power simulation results for genSKAT meta-analyses and other methods when the number of participants is the same in all 4 studies, which means different number of observations for studies with 1 observation per person and studies with 3 observations per person. GenSKAT meta-analysis assuming homogeneity has similar power as genSKAT joint analysis on pooled data. They both use the same average minor allele frequencies to calculate the weights for rare variants. When the genetic effect are homogeneous across all studies, which is scenario 1, genSKAT meta-analysis assuming homogeneity is more powerful compared to the one assuming heterogeneity. FamSKAT meta-analysis has the lowest power because it only includes 4,000 baseline observations while other methods use 8,000 observations. In scenarios 2 and 3, the difference effect sizes introduce some heterogeneity. The power of genSKAT meta-analysis assuming homogeneity and the power of genSKAT meta-analysis assuming heterogeneity are closer. If studies come from different populations, e.g. studies share a small proportion of shared causal variants, the minor allele frequencies are different, or the LD structure varies, genSKAT meta-analysis assuming homogeneity may lose more power and be out-performed by genSKAT meta-analysis assuming heterogeneity. The power of famSKAT meta-analysis assuming heterogeneity is similar
in scenarios 2 and 3, because the effect size changes happen to the same number of individuals used by this method.

The right panel from Figure 4 shows the power simulation results for genSKAT meta-analyses and other methods when the number of observations is the same for all 4 studies, arising from different number of individuals for studies with 1 observation per person and studies with 3 observations per person. As in the left panel of Figure 4, genSKAT meta-analysis assuming homogeneity and the joint genSKAT analysis have similar power. They out-perform genSKAT meta-analysis assuming heterogeneity in scenario 1. In scenario 2 and 3, the different effect sizes introduce some heterogeneity. The power difference between genSKAT meta-analyses assuming homogeneity and assuming heterogeneity are decreased. FamSKAT meta-analysis has the lowest power among all methods since it only uses part of the data. The power of famSKAT meta-analysis increases from scenario 2 to scenario 3, because fewer individuals have smaller effect sizes in the latter scenario.
Figure 4 Power comparison for genSKAT meta-analysis and other approaches

Left: same number of individuals in all studies; right: same number of observations in all studies. Scenario 1: all studies have the same effect sizes; scenario 2: studies with a single measurement per person have half of the effect sizes of studies with repeated measurements; scenario 3: studies with repeated measurements have half of the effect sizes of studies with a single measurement per person.
4.4 Discussion

We discussed gene-based meta-analysis for the genSKAT statistic. In meta-analyses, different studies may have different study designs. Some designs consist of only unrelated participants, other designs include family samples, with or without longitudinal data or repeated measurements. Ignoring the correlation structure in related individuals or multiple measures from the same person may result in elevated type-I error. Omitting correlated observations may lead to power loss. Because genSKAT is a flexible method that can adjust for the correlations between rare variants, family members, and repeated measurements, it can be applied to various study designs. In this chapter, we showed how to meta-analyze results from genSKAT assuming homogeneity or heterogeneity of genetic effects.

In our simulations, we introduced 4 study designs: family studies with or without repeated measurements, and studies of unrelated individuals with or without repeated measurements. We showed that genSKAT and famSKAT meta-analyses assuming homogeneity or heterogeneity have the correct type I error under various situations: 4 family studies with repeated measurements, 10 family studies with repeated measurements, and 4 studies with 1 from each of 4 study designs.
We performed simulation studies to evaluate power when sample sizes and effect sizes varied across studies. We simulated 6 different scenarios with different sample sizes and genetic effect sizes. When all studies shared the same effect sizes, genSKAT meta-analysis assuming homogeneity and genSKAT joint analysis on pooled individual level data had similar power in the scenarios evaluated. This is because the test statistics for both approaches are the same when using the same weights for rare variants when there are no covariates in the model. GenSKAT meta-analysis assuming homogeneity outperformed genSKAT meta-analysis assuming heterogeneity when there is no heterogeneity present. When some heterogeneity is introduced by the differences in genetic effect sizes, the power of genSKAT meta-analysis assuming homogeneity and the power of genSKAT assuming heterogeneity are closer. Because genSKAT meta-analyses take advantage of the full data, they are more powerful than famSKAT meta-analyses on baseline data in our simulations.

The genSKAT meta-analysis approach only requires genSKAT score statistic and minor allele frequency for each variant, and a covariance matrix of the scores from each study. GenSKAT meta-analysis is not based on regression coefficients because the estimated regression coefficients are usually unstable due to the low allele frequency of rare
variants. Meta-analysis does not require individual level data, thus avoiding the ethical concerns with sharing human genetic data. It also reduces the resources needed to transfer large amount of genetic data between analysts. More importantly, meta-analysis allows different studies to adjust for different variables.

With faster and cheaper genotype sequencing technologies, more genetic data will be available from existing and new studies, some with longitudinal measurements available. Meta-analysis of genSKAT results is an efficient way to increase the power to detect novel rare variant associations.

Although we performed simulations assuming repeated phenotype measurements, genSKAT can also be applied to meta-analyze multiple traits when such data are available. Lee et al have shown that Hom-Meta-SKAT-O and Het-Meta-SKAT-O are more powerful than Hom-Meta-SKAT and Het-Meta-SKAT respectively [32]. Replacing SKAT meta-analysis statistic by genSKAT meta-analysis statistic in Hom-Meta-SKAT-O and Het-Meta-SKAT-O, we can easily obtain genSKAT-O meta-analysis to further improve power.
Chapter 5 Future Work and Summary

5.1 Summary

Genetic loci identified to date by single variant association tests only account for a small proportion of the heritability for most complex traits and diseases. The unexplained heritability may be due to rare variants or gene by environment interaction. In chapter 2, we proposed a general SKAT framework to perform gene-based main effect test on family samples with multiple measurements per person. This approach is also useful to jointly analyze multiple traits, thereby improving our understanding of complex diseases. GenSKAT is an extension of SKAT and famSKAT that adjusts for the correlations between rare variants, between individuals, and between multiple measurements within a person. In chapter 3, we further discussed genSKAT for gene by environment interaction effects. Even though there are several statistical models designed for rare variant association studies, it is still difficult to find rare variant associations in real datasets. One reason is the relatively small sample size of each study. Meta-analysis is a popular approach to increase the power by combining multiple studies with the same outcome. In chapter 4, we discussed a meta-analysis framework for genSKAT results. Various simulations were performed to show that the type I error is well controlled in different scenarios. Using a different set of simulations, we also evaluated the power of genSKAT meta-analyses assuming homogeneous or heterogeneous genetic main effects across
studies, genSKAT joint meta-analysis, and famSKAT meta-analysis assuming heterogeneous genetic main effects under various sample sizes and effect sizes. Combining all these projects together, we contribute methodologies to detect rare variants associations by taking advantage of more information.

5.2 Discussion and Future Work

In chapter 2, we developed genSKAT for continuous outcomes. For categorical outcomes, the covariance structure due to familial correlation needs to be incorporated in other ways than through a kinship matrix. This is a typical limitation of genSKAT and famSKAT. The original SKAT can be applied to both continuous and categorical outcomes because it does not require adjustment for family structure. Future work on exploring ways to define the covariance matrix or finding a transformation of the outcome is needed.

We performed simulations assuming the phenotype is repeated measurements. Further simulations for jointly analyzing multiple traits is needed. Two additional covariance matrices are introduced into the genSKAT model for correlations between multiple
measurements, which means more parameters are estimated and hence a larger sample is needed. For repeated measurements, we may assume specific covariance structure based on prior information of the outcome to reduce the number of parameters. For multiple traits, it is usually hard to pre-specify the covariance structure. If we set the covariance to be unstructured, there are \( m(m + 1) \) parameters to be estimated for these two matrices. For 1 more trait, we need to additionally estimate \( 2(m + 1) \) parameters. If the sample size is small, we may see unstable estimates of the parameters. Further work on how the number of traits affects type I error and power is needed. It is also worthy to explore other ways of using fewer parameters to incorporate the correlation of multiple measurements in the models. Additionally, we may investigate the model performance using different types of covariance structures, e.g. time-dependent genetic covariance between multiple measurements.

In the process of creating genSKAT statistic, we need to perform matrix multiplication and calculate inverses of multiple matrices, inducing computational intensity for large samples with a high number of repeated measurements per person. Further work on improving computational efficiency is needed.
In chapter 3, we discussed genSKAT to test for the presence of gene by environment interaction effects. Chen et al. have proposed a joint test of both main effect and G-E interaction effect under the SKAT framework [7]. We may be able to use similar ideas to develop a joint test in the genSKAT framework.
Bibliography


8 Chen, Han, Meigs, James B, and Dupuis, Jos. Sequence kernel association test for quantitative traits in family samples. Genetic epidemiology, 37, 2 (2013), 196-204.


11 Derkach, Andriy, Lawless, Jerry F, and Sun, Lei. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic epidemiology*, 37, 1 (2013), 110-121.


24 Jiang, Duo and McPeek, Mary Sara. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genetic epidemiology*, 38, 1 (2014), 10-20.


28 Kooperberg, Charles, LeBlanc, Michael, Dai, James Y, and Rajapakse, Indika. Structures and assumptions: strategies to harness gene$\times$ gene and


37 London, Stephanie J, Gao, Wei, Gharib, Sina A et al. ADAM19 and HTR4 Variants and Pulmonary Function Cohorts for Heart and Aging Research in Genomic


56 Szumlanski, C; Otterness, D; Her C; Lee D; Brandriff B; Kelsell D; Spurr N; Lennard L; Wieben E; Weinshilboum R. Thiopurine methyltransferase pharmacogenetics: human gene cloning and characterization of a common polymorphism. *DNA and cell biology*, 15, 1 (1996), 17-30.


64 Zeggini, Eleftheria and Ioannidis, John PA. Meta-analysis in genome-wide association studies (2009).
Curriculum Vitae

WEI GAO
Department of Biostatistics, Boston University
gaowei@bu.edu

EDUCATION
Boston University, Boston, MA
Ph.D. in Biostatistics
2010~2015

The George Washington University, Washington, DC
M.S. in Biostatistics
2008~2010

Fudan University, Shanghai, China
B.S. in Mathematic
2004~2008

RESEARCH AND WORK EXPERIENCE
Department of Biostatistics, Boston University
Research Assistant
2010~2015

AstraZeneca, Waltham, MA
Intern
Summer 2013

Department of Biostatistics, Boston University, Boston, MA
Teaching Assistant
Fall 2012, Fall 2013

National Minority Quality Forum, Washington, DC
Department of Epidemiology and Biostatistics, George Washington University
Senior Research Assistant
Spring 2010

The Rock Creek Group, Washington, DC
Junior Research Assistant/Programmer
Summer 2009
PUBLICATIONS


10. Li X, Gao W. The application of the parallel multistep hybrid method to the numerical solution for a class of index-2 differential algebraic equations. *Journal of Central South
101


**PAPERS UNDER REVIEW**


14 Hansen JG, **Gao W**, Dupuis J, O'Connor GT, ..., Cassano PA. Association of 25-Hydroxyvitamin D Status and Genetic Variation in the Vitamin D Metabolic Pathway with FEV1 in FHS.

15 Araki T, Nishino M, **Gao W**, Dupuis J, Hunninghake GM, Washko GR, O'Connor GT, Hatabu H. Normal Thymus in Adults: Appearance on CT and Associations with Age, Sex, BMI and Smoking.

**ABSTRACTS AND POSTERS**

1 **Gao W**, O’Connor GT, Dupuis J. Sequence Kernel Association Test in Family Samples with Repeated Measurements or Multiple Traits. *SPH Research Day*, Boston University, Boston, MA, Nov. 2013.


**PRESENTATIONS**

2 Featured presentation at GSI Research Symposium, Boston University, Boston, MA, Nov. 2013.


HONORS

- Shanghai Outstanding College Graduates, Aug. 2008.
- Fudan University People's Scholarship, 1st Prize, 2008.
- 2nd Prize in the national round, 1st Prize in Shanghai, 10th “Challenge Cup” National Competition of Chinese College Students’ Extracurricular Academic and Scientific Achievements, Nov. 2007.