

2018

# Evaluation of marker density for population stratification adjustment and of a family-informed phenotype imputation method for single variant and variant-set tests of association

---

<https://hdl.handle.net/2144/33081>

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

BOSTON UNIVERSITY  
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**EVALUATION OF MARKER DENSITY FOR POPULATION STRATIFICATION  
ADJUSTMENT AND OF A FAMILY-INFORMED PHENOTYPE IMPUTATION  
METHOD FOR SINGLE VARIANT AND VARIANT-SET TESTS OF  
ASSOCIATION**

by

**YUNING CHEN**

B.S., Beijing Institute of Technology, 2011  
M.S., Marquette University, 2013

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

2018



Approved by

First Reader

---

Josée Dupuis, Ph.D.  
Professor of Biostatistics

Second Reader

---

Gina M. Peloso, Ph.D.  
Assistant Professor of Biostatistics

Third Reader

---

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

## Acknowledgments

First, I would like to express my deepest gratitude to my major advisor Dr. Josée Dupuis, who has helped me, supported me and changed my life. Josée not only teaches me the knowledge of statistics and genetics, but also shows me how to be a good researcher and a good person. Josée's advices and guidance help me move forward to get my PhD. Without Josée, I would not have been able to finish this dissertation. I would also like to sincerely thank my whole dissertation committee: Dr. Gina Peloso, Dr. Ching-Ti Liu, Dr. Anita DeStefano and Dr. Kathryn Lunetta for their very helpful and invaluable advice for my research.

I would like to thank Shuai for her tremendous help in my first two years. My friends Yicheng and Qiang have been very supportive and I wish all of my friends a brighter future.

Finally, I want to express my gratitude to my parents for their love and support through my journey towards this degree.

**EVALUATION OF MARKER DENSITY FOR POPULATION STRATIFICATION  
ADJUSTMENT AND OF A FAMILY-INFORMED PHENOTYPE IMPUTATION  
METHOD FOR SINGLE VARIANT AND VARIANT-SET TESTS OF  
ASSOCIATION**

**YUNING CHEN**

Boston University Graduate School of Arts and Sciences, 2018

Major Professor: Josée Dupuis, Professor of Biostatistics

**ABSTRACT**

Whole exome sequencing (WES) data cover only 1% of the genome and is designed to capture variants in coding regions of genes. When associating genetic variations with an outcome, there are multiple issues that could affect the association test results. This dissertation will explore two of these issues: population stratification and missing data. Population stratification may cause spurious association in analysis using WES data, an issue also encountered in genome-wide association studies (GWAS) using genotyping array data. Population stratification adjustments have been well studied with array-based genotypes but need to be evaluated in the context of WES genotypes where a smaller portion of the genome is covered. Secondly, sample size is a major component of statistical power, which can be reduced by missingness in phenotypic data. While some phenotypes are hard to collect due to cost and loss to follow-up, correlated phenotypes

that are easily collected and are complete can be leveraged in tests of association.

First, we compare the performance of GWAS and WES markers for population stratification adjustments in tests of association. We evaluate two established approaches: principal components (PCs) and mixed effects models. Our results illustrate that WES markers are sufficient to correct for population stratification. Next, we develop a family-informed phenotype imputation method that incorporates information contained in family structure and correlated phenotypes. Our method has higher imputation accuracy than methods that do not use family members and can help improve power while achieving the correct type-I error rate. Finally, we extend the family-informed phenotype imputation method to variant-set tests. Single variant tests do not have enough power to identify rare variants with small effect sizes. Variant-set association tests have been proven to be a powerful alternative approach to detect associations with rare variants. We derive a theoretical statistical power approximation for both burden tests and Sequence Kernel Association Test (SKAT) and investigate situations where our imputation approach can improve power in association tests.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Overview . . . . .	1
1.2	Population Stratification . . . . .	3
1.2.1	Principal Component Analysis . . . . .	3
1.2.2	Mixed Effects Model . . . . .	4
1.3	Association Test . . . . .	6
1.3.1	Single-Variant Analysis . . . . .	6
1.3.2	Variant-set Test . . . . .	7
1.4	Dissertation Outline . . . . .	9
<b>2</b>	<b>Evaluation of Marker Density for Population Stratification Adjustment</b>	<b>11</b>
2.1	Introduction . . . . .	11
2.2	Methods . . . . .	14
2.2.1	Genotypes . . . . .	14
2.2.2	PCs and GRM Computation . . . . .	15
2.2.3	Simulation Study Subjects . . . . .	15

2.2.4	Comparison of PCs and GRM Generated Using GWAS and WES Variants . . . . .	16
2.2.5	Comparison of Genomic Control Factor and Type-I Error Rate . . .	18
2.2.6	Power Evaluation . . . . .	19
2.2.7	Comparison Using FHS GWAS and Exome Chip Data . . . . .	20
2.3	Results . . . . .	22
2.4	Discussion . . . . .	32
<b>3</b>	<b>A Family-Informed Phenotype Imputation Approach</b>	<b>36</b>
3.1	Introduction . . . . .	36
3.2	Methods . . . . .	38
3.2.1	Phenotype Imputation for Family Data . . . . .	38
3.2.2	Phenotype Imputation for Population-Based Studies . . . . .	40
3.2.3	Relationship with Imputation Using Regression Model . . . . .	41
3.2.4	Power of Single Variant Test . . . . .	43
3.2.5	Analysis with Combined Observed and Imputed Data . . . . .	44
3.2.6	Strategy to Analyze the Observed and Imputed Data . . . . .	46
3.2.7	Factors Influencing Imputation Accuracy . . . . .	47
3.2.8	Simulation Evaluation for Single Variant Test . . . . .	48
3.2.9	Application of 2 Hour Glucose in FHS . . . . .	51
3.3	Results . . . . .	52
3.4	Discussion . . . . .	64

<b>4</b>	<b>Extension of a Phenotype Imputation Approach to Variant-Set Tests</b>	<b>67</b>
4.1	Introduction . . . . .	67
4.2	Methods . . . . .	68
4.2.1	Power of Burden Test . . . . .	68
4.2.2	Power of SKAT . . . . .	71
4.2.3	Simulation . . . . .	74
4.2.4	Application of 2 Hour Glucose in FHS . . . . .	76
4.3	Results . . . . .	77
4.4	Discussion . . . . .	85
<b>5</b>	<b>Summary and Future Work</b>	<b>88</b>
	<b>Bibliography</b>	<b>90</b>
	<b>Curriculum Vitae</b>	<b>94</b>

# List of Tables

2.1	Origin of 1000G EA and AA founders . . . . .	22
2.2	Genomic control factor $\lambda_{GC}$ from simulation studies . . . . .	26
2.3	Relative type-I error rate in simulation studies . . . . .	28
2.4	Power from simulation studies in EA + AA samples . . . . .	30
2.5	Power from simulation studies in EA samples . . . . .	31
2.6	Association test results from FHS . . . . .	32
3.1	Power evaluation results of simulated data with 20% missing percentage . . . . .	55
3.2	Power evaluation results of simulated data with 50% missing percentage . . . . .	56
3.3	Power evaluation results of simulated data with 80% missing percentage . . . . .	57
3.4	Relative Type-I error rate of family data . . . . .	58
3.5	Power evaluation for the combined observed and imputed family data . . . . .	59
3.6	Imputation accuracy of family data (2 parents + 2 offsprings) . . . . .	60
3.7	Imputation accuracy of family data (2 parents + 4 offsprings) . . . . .	60
3.8	Association tests results of FG and 2 hour glucose in FHS . . . . .	61
3.9	Median P-values from FHS 2 hour glucose data . . . . .	63

4.1	Power of combined observed and imputed data for unrelated samples with 20% missing percentage in variant-set tests. . . . .	77
4.2	Power of combined observed and imputed data for unrelated samples with 50% missing percentage in variant-set tests. . . . .	78
4.3	Relative Type-I error rate of unrelated data in variant-set tests. . . . .	79
4.4	Relative Type-I error rate of family data in variant-set tests. . . . .	79
4.5	Power of unrelated data (Imp+Obs) with 20% missing percentage in variant-set tests . . . . .	81
4.6	Power of unrelated data (Imp+Obs) with 50% missing percentage in variant-set tests . . . . .	82
4.7	SKAT Power of family data (Imp+Obs) with 20% missing percentage in variant-set tests . . . . .	83
4.8	SKAT Power of family data (Imp+Obs) with 50% missing percentage in variant-set tests . . . . .	84
4.9	Variant-set association tests results of FG and 2 hour glucose in FHS . . . . .	85
4.10	Median P-values from FHS 2 hour glucose data in variant-set tests . . . . .	85

# List of Figures

2.1	Population stratification in 1000G Phase 3 data. . . . .	17
2.2	Pair-wise comparison of kinship coefficients computed using GWAS and WES variants . . . . .	24

# List of Abbreviations

AA	African Ancestry
AIMs	Ancestry-Informative Markers
BN	Balding-Nichols
CHARGE	Cohorts for Heart and Aging Research in Genomic Epidemiology
CVD	Cardiovascular Disease
EA	European Ancestry
FG	Fasting Glucose
FHS	Framingham Heart Study
GLMM	Generalized Linear Mixed Effects Model
GLS	Generalized Least Squares
GRM	Genetic Relationship Matrix
GWAS	Genome-Wide Association Studies
IBS	Identical-By-State
LD	Linkage Disequilibrium
LMM	Linear Mixed Effects Model
MAF	Minor Allele Frequency
MAR	Missing At Random
MCAR	Missing Completely At Random
MLE	Maximum Likelihood Estimator
MNAR	Missing Not At Random

MSE	Mean Square Error
MVN	Multivariate Normal
NCP	Non-Centrality Parameter
NMM	Noise Measurement Model
PC	Principal Component
PCA	Principal Component Analysis
PheWAS	Phenome-Wide Association Studies
SHARe	SNP Health Association Resource
SKAT	Sequencing Kernel Association Test
SNVs	Single Nucleotide Variants
SVD	Singular Value Decomposition
VB	Variational Bayes
WES	Whole Exome Sequencing

## Chapter 1 Introduction

### 1.1 Overview

Genome-wide association studies (GWAS) have identified thousands of single nucleotide variants (SNVs) associated with complex traits and phenotypes [1, 2, 3]. GWAS include SNVs covering the human genome, both within and outside of genes and regulatory regions. However, associated variants often fall outside of genes and regulatory regions, and do not contribute to further our understanding of the genetic architecture of a disease because of our current, limited understanding of the function of the majority of the genome. In order to identify causal variants, whole exome sequencing (WES) has become one of the leading strategies in association studies because of the easy interpretability of WES variants [10]. Exons provide a good source of variants potentially influencing complex traits and diseases even though they only cover about 1% of the whole genome.

With the increase of collaboration among researchers all over the world, it is possible to collect data on individuals from different ancestry so that the sample size of a study can be maximized and statistical power can be increased. However, GWAS still suffer from two issues that could affect the association test results: population stratification and missing data. Population stratification, which is the allele frequency difference between cases and controls due to ancestry difference, is a source of inflated type-I error when it is not corrected [4]. Missing data are common in epidemiology studies due to cost and lost to

follow-up.

In any genetic study, information about ancestry is not always available or collected. In addition, the collected self-reported ancestry information may not be accurate or too broad. Including samples from different ancestries can introduce population stratification, which can cause false-positive results. Among all methods to account for population stratification, principal component analysis (PCA) and mixed effects models are the most popular approaches because of their convenience and good performance in GWAS.

While many approaches have been developed for genotype imputation, little attention has been given to phenotype imputation. The most common way to handle missing phenotype data is remove the individuals with missing observations, which decreases the sample size and the statistical power to detect the association, and could introduce bias if the missing mechanism is not missing completely at random (MCAR) or missing at random (MAR). Some phenotypes are hard to collect but related phenotypes may be available and can be exploited. For example, a CT scan is needed to measure visceral fat. However, one can use waist hip ratio or waist circumference as alternatives, which only requires a tape measure. Information contained in the related phenotypes can be included in the phenotype imputation.

In this dissertation, we focus on population stratification and missing data issues in GWAS

and exome-wide association studies. We investigate the performance of WES variants on population stratification adjustment approaches and compare the results from association tests using GWAS array and WES variants to correct for population stratification. We also develop a phenotype imputation approach that can include information contained in related phenotypes and family structure. Lastly, we extend the phenotype imputation method from single variant tests to variant-set tests, and investigate its performance under different situations through extensive simulation work.

## **1.2 Population Stratification**

### **1.2.1 Principal Component Analysis**

Population stratification, the allele frequency difference between cases and controls due to ancestry, can be corrected by PCA [4, 5]. PCA is a dimension reduction method that converts possibly correlated observations into a set of linearly independent variables: the principal components (PCs). In genetic studies, PCA can be applied to genotype data to reduce the dimension of the data and the population stratification can be captured by the PCs. Because the top PCs can explain most of the variability in ancestry, they can be included as covariates in association tests to correct for population stratification. Models without PC adjustment can result in the identification of spurious association in the presence of population stratification.

PCA starts with computing the covariance matrix of the genotype data, and then the

eigenvectors on this matrix are obtained. Alternatively, one can also perform singular value decomposition (SVD) on the genotype matrix. The  $k$ th eigenvector or PC corresponds to the  $k$ th eigenvalue  $\lambda_k$ . Each eigenvalue is proportional to the percentage of variance explained by the corresponding eigenvector. The first PC can explain the largest variance of the genotypic data, and each succeeding PC has the highest variance under the constraint that it is orthogonal to the preceding PCs.

Mathematically, PCs are linear combinations of the genotype vectors. Genetic variants with larger weight in the linear combination show bigger difference between populations and hence contribute more in the PC computation than variants with a smaller weight.

These variants are also referred to as ancestry-informative markers (AIMs). In this dissertation, the PC computation is done using the smartpca program in the EIGENSOFT package [5].

### 1.2.2 Mixed Effects Model

A linear mixed effects model can be used to account for population stratification when the phenotype is continuous [7]. The model can be written as  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ , where  $\mathbf{y}$  is the  $n \times 1$  phenotype vector,  $\mathbf{X}$  is the  $n \times q$  matrix including covariates and the genotype of a single variant,  $\boldsymbol{\beta}$  is the  $q \times 1$  vector of the regression coefficients of fixed effects,  $\mathbf{Z}$  is the  $n \times t$  incidence matrix relating each item in the phenotype vector to one of the  $t$

subgroups,  $\mathbf{u}$  is the  $t \times 1$  random effect vector and  $\mathbf{e}$  is the  $n \times 1$  vector of residuals.

We further assume that  $\text{Var}(\mathbf{u}) = \sigma_g^2 \mathbf{K}$  and  $\text{Var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}$ , where  $\sigma_g^2$  and  $\sigma_e^2$  are the polygenic and environmental variance components. Random effect  $\mathbf{u}$  is independent of the residuals  $\mathbf{e}$  and  $\mathbf{K}$  is the kinship matrix describing the relationship between the  $t$  subgroups. In GWAS, we usually assume that the random effect is the polygenic effect for each sample, which implies that  $t = n$  and  $\mathbf{Z} = \mathbf{I}$ . The total phenotypic variance can be written as  $\text{Var}(\mathbf{y}) = \Sigma = \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}$ .

The parameters  $\sigma_g^2$  and  $\sigma_e^2$  are called the variance components, and are assumed to be unknown. To solve the mixed model equation, the variance components must first be estimated. Once this is done, a generalized least squares (GLS) procedure may be used to estimate  $\beta$  by  $\hat{\beta} = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma^{-1} \mathbf{y}$ .

When using linear mixed effects models to account for population stratification, the kinship matrix  $\mathbf{K}$  is computed from the genotype data. It is an  $n \times n$  matrix with each entry describing the relationship between each pair of individuals. In studies with known family structure,  $\mathbf{K}$  can also be computed from pedigree. When family structure is unknown, there are two types of commonly-used kinship matrix: identical-by-state (IBS) and Balding-Nichols (BN) kinship matrix [7]. IBS kinship matrix measures the proportion of alleles IBS between each pair of individuals. In the BN kinship matrix, the genetic

relationship between individuals  $i$  and  $j$  is estimated through

$$\sigma_{ij} = \frac{1}{m} \sum_{k=1}^m \frac{(x_{k,i} - 2\hat{p}_k)(x_{k,j} - 2\hat{p}_k)}{2\hat{p}_k(1 - \hat{p}_k)} \quad (1.1)$$

where  $m$  is the total number of variants,  $x_{k,i}$  is the genotype of the  $i$ th individual on the  $k$ th variant and  $\hat{p}_k$  is the minor allele frequency (MAF) of the  $k$ th variant.

### 1.3 Association Test

#### 1.3.1 Single-Variant Analysis

The standard GWAS framework includes one phenotype and association between the phenotype and the SNVs is evaluated one SNV at a time using logistic (binary trait) or linear (continuous trait) regression. The statistical model for the phenotype value of the  $i$ th individual  $y_i$  is  $f(y_i) = \beta_0 + \beta_1 g_i + \gamma_1 \mathbf{x}_i$ , where  $f(y_i) = y_i$  for a quantitative phenotype and  $f(y_i) = \text{logit } P(y_i = 1)$  for a binary trait. A normally-distributed error term  $e_i$  with mean zero and variance  $\sigma^2$  is added to the model for a quantitative phenotype.  $\beta_1$  is the regression coefficient for the genotype  $g_i$  and  $\gamma_1$  is the vector of regression coefficients for the covariates vector  $\mathbf{x}_i$ . To test the association between  $g_i$  and  $f(y_i)$ , a hypothesis test is performed with the null  $H_0 : \beta_1 = 0$ .

To simplify the linear regression model, we can regress  $y_i$  on  $\mathbf{x}_i$ , obtain the residuals and use the residuals as outcome in the association test of the single variant. The model then becomes,  $r_i = \beta_0 + \beta_1 g_i + e_i$ , where  $r_i$  is the residual and  $e_i$  is the error term. The

maximum likelihood estimates are  $\hat{\beta}_0 = \frac{1}{n} \mathbf{1}^T \mathbf{r}$  and  $\hat{\beta}_1 = \frac{\mathbf{g}^T \mathbf{r}}{\mathbf{g}^T \mathbf{g}}$ , where  $\mathbf{r}$  and  $\mathbf{g}$  are the outcome and genotype vectors of all individuals and  $\mathbf{1}$  is a vector of all ones. The estimated standard error is  $\hat{\sigma} = \sqrt{\frac{\hat{\mathbf{e}}^T \hat{\mathbf{e}}}{n-2}}$ , where  $\hat{\mathbf{e}} = \mathbf{r} - \hat{\beta}_0 \mathbf{1} - \hat{\beta}_1 \mathbf{g}$ . The test statistic can then be written as  $t = \frac{\hat{\beta}_1}{\hat{\sigma}} \sqrt{\mathbf{g}^T \mathbf{g}}$ . Note that  $\mathbf{g}^T \mathbf{g}$  can be replaced by its estimation  $2np(1-p)$ , where  $p$  is the MAF of the single variant. We can further standardize  $g_i$  so that  $\mathbf{g}^T \mathbf{g} = n$ . Under the alternative hypothesis of association, the statistic  $t$  follows a normal distribution with non-centrality parameter  $\frac{\beta_1}{\sigma} \sqrt{n}$  and variance 1 [37].

Statistical power is the probability of correctly rejecting the null hypothesis. In GWAS, it is equivalent to the probability of detecting a true association. Based on the derivation of the non-centrality parameter(NCP) above, we can estimate power as

$$\Phi\left(\Phi^{-1}(\alpha/2) - \frac{\beta_1}{\sigma} \sqrt{n}\right) + 1 - \Phi\left(\Phi^{-1}(1 - \alpha/2) - \frac{\beta_1}{\sigma} \sqrt{n}\right) \quad (1.2)$$

where  $\Phi$  is the cumulative density function of the standard normal distribution and  $\alpha$  is the significance level in GWAS.

### 1.3.2 Variant-set Test

Even though GWAS have been very successful in detecting common variants associated with complex traits or phenotypes, much of the heritability is still unexplained and few rare genetic variants have been found to be associated with diseases thus far [27]. Single variant tests do not have enough power to detect the association with rare variants, hence

aggregating information from multiple rare genetic variants is an alternative in association testing to improve power.

One class of aggregation tests is the burden test, which collapses the genotypes of rare variants into a burden score and assesses the association between the burden score and the phenotype. Similar to the single-variant analysis, we can write the statistical model for the multiple variant test as, for the  $i$ th individual,  $f(y_i) = \beta_0 + \beta \mathbf{G}_i + \gamma \mathbf{x}_i$ , where  $\gamma$  is the vector of regression coefficients for the vector of covariates  $\mathbf{x}_i$  and  $\beta$  is the vector of regression coefficients for the genotype vector  $\mathbf{G}_i$ . The burden score can be written as  $C_i = \sum_{j=1}^p w_j G_{ij}$ , where  $w_j$  is the weight for variant  $j$ . It is equivalent to testing  $H_0 : \beta_c = 0$  in  $f(y_i) = \beta_0 + \beta_c C_i + \gamma \mathbf{x}_i$ .

Burden tests have the highest power when all variants in the same region have the same direction of effect, but often most variants in the region have little or no effect on the phenotype. In addition, there could be both protective and deleterious variants in the region. To allow for the presence of both deleterious and protective variants, a variance component test, sequence kernel association test (SKAT), was developed in a multiple regression framework and uses a variance-component score test which is flexible and computationally efficient.

SKAT is based on the same regression model as in the burden test with  $\beta = [\beta_1, \dots, \beta_p]$ ,

defined as the effect sizes of the  $p$  variants on the phenotype. The SKAT statistic can be written as  $Q = (y - \bar{y})^T K (y - \bar{y})$ , where  $K = GWG^T$ ,  $G$  is the genotype matrix and  $W = \text{diag}(w_1, \dots, w_p)$  is the matrix of weights. Wu et al. [27] suggested to use  $\sqrt{w_j} = \text{Beta}(\text{MAF}_j, 1, 25)$  so that rare variants have higher weights and variants with MAF between 1% and 5% also have reasonable nonzero weights.

One big advantage of SKAT is that its test statistic can be computed using the individual variant test statistics by  $Q = \sum_{j=1}^p w_j S_j^2$ , where  $S_j$  is the individual score statistic for testing the association between the phenotype and the individual variant  $j$ . In addition, SKAT uses a score test, which only needs to fit the null model once and hence it can be computationally efficient.

#### **1.4 Dissertation Outline**

In the first project, we compare the performance of GWAS and WES variants on population stratification adjustment through simulation. Specifically, we compare the PCs, kinship coefficients in the genetic relationship matrix (GRM) computed using GWAS and WES variants, and association results between the PC-adjusted model and the mixed effects model in terms of genomic control factor, type-I error rate and statistical power.

In the second project, we develop a family-informed phenotype imputation approach that incorporates the information contained in family structure and additional correlated

phenotypes. We show that taking family structure into consideration when imputing can improve imputation accuracy. Simulation studies show that our imputation approach can improve the statistical power to detect the association while achieving the correct type-I error rate. In addition, we also derive the approximated NCP in association tests of the combined observed and imputed phenotype so that it is possible to compute the theoretical power for the association testing in the study design phase. We show that the theoretical power from our NCP derivation is very close to the empirical power from simulations, and investigate the situations where our method can improve power.

In the third project, we extend the phenotype imputation approach to variant-set tests. We focus on two commonly-used variant-set tests: burden test and SKAT. We first derive the approach to compute the theoretical power for these two tests, and then verify our derivation through simulation studies. We examine the type-I error rate and power of jointly analyzing observed and imputed phenotype under different conditions through extensive simulations. We also validate our findings using a real dataset from the Framingham Heart Study (FHS).

## **Chapter 2 Evaluation of Marker Density for Population Stratification Adjustment**

### **2.1 Introduction**

Genome-wide association studies (GWAS) have been proven to be a useful tool to discover single nucleotide variants (SNVs) associated with complex traits [1, 2, 3].

Typically, GWAS are restricted to study subjects that share the same ancestry. Population stratification, which is the allele frequency difference between cases and controls due to ancestry difference, occurs when there are multiple population groups within a sample.

The association test results can be affected by population stratification, which can result in an inflated type-I error rate. Current methods for correcting population stratification include principal components (PCs) of the genotypes [4, 5], genomic control factor [6], linear mixed effects model (LMM) and generalized linear mixed effects model (GLMM) using an empirical kinship matrix [7, 8] and structured association [9].

PC correction has been widely used in GWAS. Population stratification can be corrected by including genetic PCs as covariates in a linear regression model for continuous traits or a logistic regression model for binary traits. In contrast, in the genomic control method, an overall inflation factor is used to adjust the association test statistic at every marker. Some markers have a bigger difference in allele frequencies across different populations, while some markers are less affected by population stratification. The overall inflation factor treats all markers the same and hence it may over-adjust markers with small differentiation across ancestral populations and under-adjust markers with strong differentiation. Yet another approach, LMMs, utilize a variance component method to model genetic

relationships. The model includes an empirically estimated genetic relationship matrix (GRM), which takes advantage of the high-density genotype information and estimates the variance parameters under the null model assuming the effect of any given marker on the phenotype is very small. The SNV effect is modeled as a fixed effect and a random intercept is included to model the relatedness among study subjects. Lastly, the structured association method adjusts for population stratification by assigning samples to subpopulation clusters and combines the association results of each cluster. This approach is highly sensitive to the number of subpopulation clusters and has intensive computational cost for large data sets, such as GWAS.

These population stratification adjustment approaches are developed in the context of GWAS, which include common markers across the whole genome. However, the performance of these methods in association analyses has not been evaluated in studies with WES data. Whether WES genotypes are sufficient to appropriately model population stratification is an emerging question in studies without GWAS genotypes, when PCs and GRM can only be computed using WES markers. Belkadi et al. [10] found a strong correlation between PCs computed using GWAS and WES variants, and that an accurate estimation of population stratification can be obtained using high-quality WES variants with  $MAF > 2\%$ . Gazal et al. [11] evaluated the performance of linkage analysis using GWAS and WES variants and showed that they had similar performance in excluding genomic regions with false-positive candidate causal variants. Smith et al. [12] also

demonstrated that accurate genetic linkage mapping can be performed using WES data. Kancheva et al. [13] showed that WES variants can provide high specificity and sensitivity for the detection of homozygous regions in consanguineous families when using GWAS variants as reference. Eu-ahsunthornwattana et al. [14] compared kinship matrices computed using different number of SNVs and different softwares, and found that the kinship coefficients computed using  $\sim 50,000$  SNVs were highly correlated with those computed using  $\sim 545,000$  SNVs.

In this chapter, we focus on the PC-based and LMM/GLMM-based population stratification adjustment methods. There are two concerns with using WES-computed PCs and GRM: 1) The PCs and GRM can only capture the genetic information in exons, while PCs and GRM computed using GWAS markers are able to capture the genetic information contained on the whole genome, hence they should be more accurate than WES-computed PCs and GRM; and 2) because the number of markers in WES is usually smaller than that in GWAS, the WES-computed PCs and GRM may contain less information than GWAS-computed PCs and GRM due to the inclusion of fewer markers in the computation. It seems obvious that the potential loss of both quality and quantity in WES variants can lead to insufficient adjustment for population stratification.

Our goals are to compare the PCs and GRM computed using GWAS and WES variants, to examine the effect of these two sets of PCs and GRM on association analysis results, and

to evaluate performance of PC-based and LMM/GLMM-based methods. We use simulations to compare the PCs and GRM computed using GWAS or WES variants in terms of genomic control factor, type-I error rate and power in association analyses. A real data set from the Framingham Heart Study (FHS) is used to compare the association results between height and a SNV in the *LCT* gene, an association known to be due to population stratification [17].

## 2.2 Methods

### 2.2.1 Genotypes

We use the genotypes from the 1000 Genome Project Phase 3 [15] dataset. GWAS variants are selected based on Illumina HumanHap300K BeadChip which is designed using international HapMap Project [19] data of individuals from CEU [20]. WES markers are annotated using the EPIACTS [21] Version 3.2.6 annotation function based on GENCODE V7 transcripts. We first apply the following quality control (QC) filters on all selected GWAS and WES SNVs in the 1000G dataset:  $MAF \geq 1\%$ , call rate  $\geq 99\%$ , Hardy-Weinberg Equilibrium P-value  $> 0.0001$ . Variants passing these QC filters are used for evaluation of the type-I error rate. We then select a subset of these SNVs for PCs and GRM computation based on the additional QC criterion: minor allele frequency  $\geq 5\%$  and only one SNV of each pair of SNVs with LD  $r^2 > 0.5$  in a 50 SNVs window.

### **2.2.2 PCs and GRM Computation**

PCs and two types of GRM, IBS and BN kinship matrix [14] are computed using three sets of variants: 1) GWAS variants; 2) WES variants; and 3) a randomly selected subset of GWAS variants that has the same number of variants as the WES set. The inclusion of the third set of variants above is to eliminate the difference in the number of variants used in PCs and GRM calculation. All PCs and GRM are computed using EIGENSTRAT [4] and EMMAX [7], respectively.

### **2.2.3 Simulation Study Subjects**

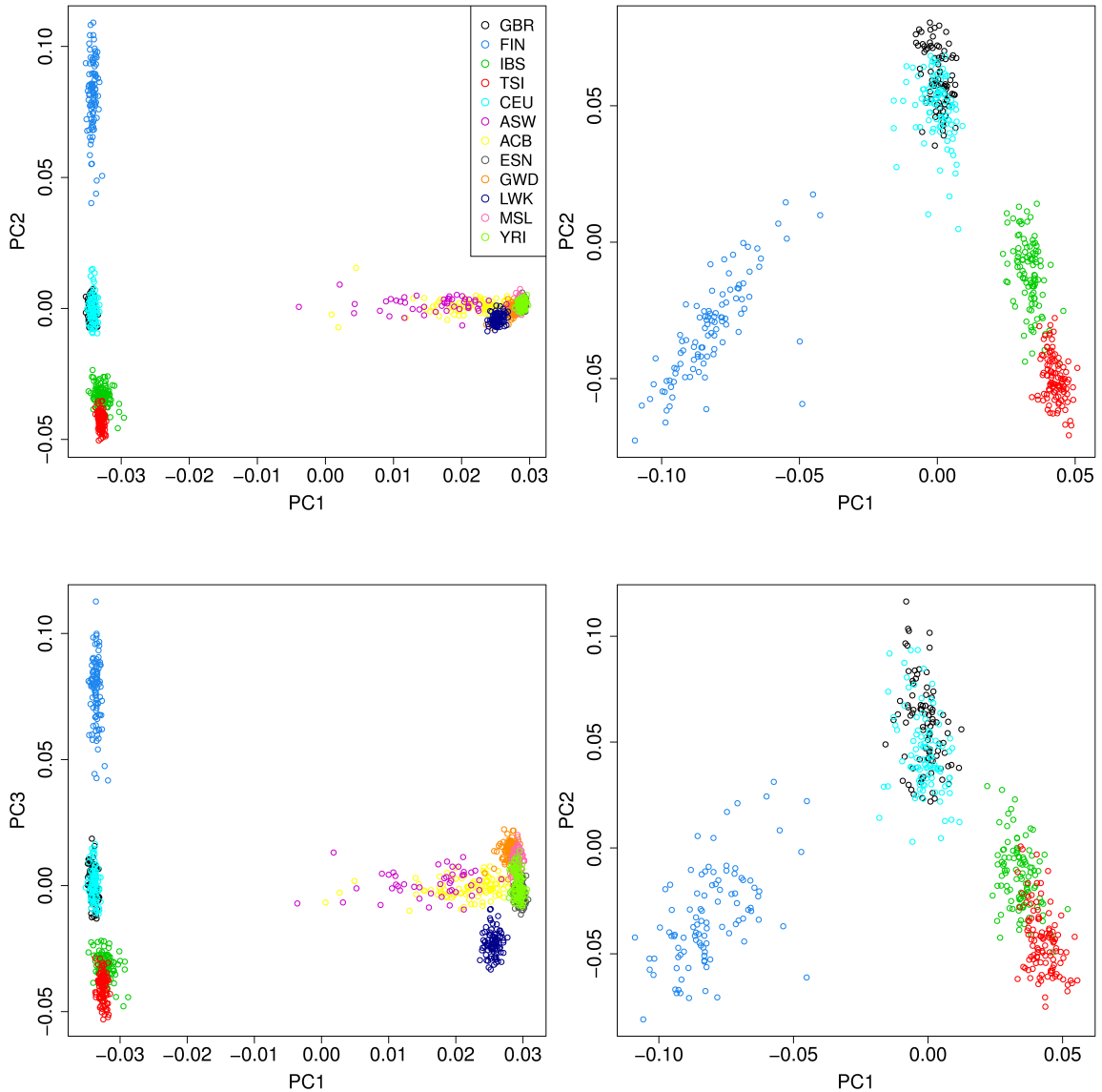
We select founders from the 1000G Phase 3 data set to generate simulated data. European ancestry (EA) founders are from 5 cohorts: FIN (Finnish in Finland), CEU (Utah Residents with Northern and Western Ancestry), GBR (British in England and Scotland), IBS (Iberian Population in Spain) and TSI (Toscani in Italy). African ancestry (AA) founders are from 7 cohorts: ASW (Americans of African Ancestry in SW USA), ACB (African Caribbeans in Barbados), LWK (Luhya in Webuye, Kenya), ESN (Esan in Nigeria), GWD (Gambian in Western Divisions in the Gambia), MSL (Mende in Sierra Leone) and YRI (Yoruba in Ibadan, Nigeria). We then compute PCs on two sets of study subjects: 1) using both EA and AA founders; and 2) restricting our analysis to EA founders only to generate a dataset with more subtle ancestry differences between sub-populations. Based on the clustering pattern (Figure 2.1), we divide the combined EA and AA 1000G founders into 6 groups (group 1, FIN; group 2, CEU, GBR; group 3, IBS,

TSI; group 4, ASW, ACB; group 5, LWK; group 6: ESN, GWD, MSI, YRI), while we do not combine cohorts in the analysis restricted to EA founders, so we simply refer to FIN, CEU, GBR, IBS and TSI as groups 1-5, respectively. Next, within each group, we generate genotypes for 333 simulated individuals for the combined EA and AA analysis and 400 for EA only analysis while maintaining the same linkage disequilibrium (LD) pattern as observed in each population between variants using the software Hapgen2 [16]. In total, 1998 and 2000 individuals are generated for the EA and AA, and EA only analyses, respectively.

#### **2.2.4 Comparison of PCs and GRM Generated Using GWAS and WES Variants**

To compare PCs, we first consider a plane developed by the first 2 PCs as they explain most of the variance. A centroid is defined by the mean vector of the first 2 PCs for each sub-group. Then the Euclidean distance from the group-specific centroid to each simulated sample is computed and a Wilcoxon signed-rank test is performed to compare the distance difference computed using GWAS variants, WES variants and the subset of GWAS variants which has the same number of variants as the WES set. In order to compare GRMs, we compute the Pearson correlation between the kinship coefficients contained in the GRM.

Figure 2.1: Population stratification in 1000G Phase 3 data.



Top left panel: scatterplot of PC1 vs. PC2 computed using GWAS markers in the combined EA and AA analysis. Top right panel: scatterplot of PC1 vs. PC2 computed using GWAS markers in the EA only analysis. Bottom left panel: scatterplot of PC1 vs. PC3 computed using WES markers in the combined EA and AA analysis. Bottom right panel: scatterplot of PC1 vs. PC2 computed using WES markers in the EA only analysis. The grouping in EA and AA founders is: group 1, FIN; group 2, CEU, GBR; group 3, IBS, TSI; group 4, ASW, ACB; group 5, LWK; group 6: ESN, GWD, MSI, YRI. The grouping in EA samples is: group 1, FIN; group 2, CEU; group3, GBR; group 4, IBS; group 5, TSI.

### 2.2.5 Comparison of Genomic Control Factor and Type-I Error Rate

To examine genomic control factor,  $\lambda_{GC}$ , defined as the ratio of the median observed test statistic to the expected test statistic under the null hypothesis, and type-I error rate, the continuous phenotype values are assigned based on a pre-specified group-specific mean level. A random error term which follows a normal distribution is then added to the assigned mean levels. In the test of a binary outcome, a total of 999 or 1000 randomly selected cases are generated for EA and AA, and EA only data, respectively. The prevalences of the simulated disease are set to 2%, 5%, 8%, 10%, 15% and 18% for the 6 EA and AA groups, while they are 2%, 5%, 8%, 10% and 15% for the 5 EA groups. The continuous and binary phenotypes are not associated with any simulated SNVs. We perform association analyses on all GWAS and WES variants with  $MAF \geq 1\%$  in spite of whether or not they are included in the PC calculation to mimic the GWAS in practice. Variants with  $MAF < 1\%$  are not included due to low statistical power to detect association with low-frequency SNVs in actual GWAS. PCs and GRM computed using GWAS and WES markers are used for population stratification adjustment.

Four linear/logistic regression models are performed: 1)  $Y \sim SNV$ , an unadjusted model; 2)  $Y \sim SNV + Group$ , a model adjusted for the true grouping assignment, which is used as the gold standard; 3)  $Y \sim SNV + PC_{GWAS}$ , a model adjusted for first 10 PCs computed from GWAS variants; and 4)  $Y \sim SNV + PC_{WES}$ , a model adjusted for the first 10 PCs computed from WES variants. Besides these four models, two LMMs for continuous trait

and two GLMMs for binary trait, using GWAS-computed and WES-computed empirical kinship matrix to account for population stratification, respectively, are also performed. In addition, for binary traits, we also include PCs in the GLMM-based method to evaluate the performance of using both PCs and GRM to adjust for population stratification. The association analyses are performed using PLINK [22] for PC-adjusted models, EMMAX [7] and the R package GENESIS [8] for mixed effects models of continuous and binary traits, respectively, and repeated 500 times.

### **2.2.6 Power Evaluation**

To evaluate power, we select 10 SNVs to be associated with the phenotype. Five of the 10 SNVs (rs2239923, rs17639812, rs3739555, rs4556520 and rs2124147 for EA and AA analysis; rs3783501, rs3741190, rs9559516, rs2071593 and rs764231 for EA only analysis) are confounded by population stratification, as indicated with a high PC weight for PC1 or PC2, and the other 5 SNVs (rs4736111, rs2238740, rs7221974, rs2071624 and rs793878 for EA and AA analysis; rs1047406, rs2276232, rs3745009, rs161557 and rs3746619 for EA only analysis) are not confounded, with low PC weights for PC1 and PC2. To select the 10 SNVs, we first rank the absolute value of the weights of the first 2 PCs in GWAS and WES sets separately for each SNV present in both GWAS and WES set. Then the ranks are added up and the top 5 SNVs (high weight, confounded by population stratification) and the last 5 SNVs (low weight, not confounded by population stratification) are selected to generate simulated phenotypes. By assuming the percentage

of the phenotypic variance explained by the SNV as  $R^2 = 1\%$ , the effect size  $\beta$  is set to be equal in all studies and it is computed using  $\beta = \sqrt{\frac{R^2}{2p(1-p)}}$ , where  $p$  is the MAF obtained in all studies. A normally distributed random error term is added to the linear combination of  $\beta$ 's and genotypes, with a group-specific mean and variance 1. The associated binary trait is then generated by assigning samples whose continuous trait value is above the 90% percentile of all samples as cases to achieve a 10% population prevalence. We evaluate the same models used in type-I error rate evaluation at  $\alpha$  level of  $10^{-4}$ . The association analyses are performed using PLINK [22] for PC-adjusted models, EMMAX [7] and the R package GENESIS [8] for mixed effects models of continuous and binary traits, respectively, and repeated 500 times.

### **2.2.7 Comparison Using FHS GWAS and Exome Chip Data**

In 1948, FHS enrolled its first participants, the Original Cohort with 5,209 individuals, from Framingham, MA. These participants aged between 30 and 62 underwent detailed physical examination, lifestyle interviews and laboratory tests every two years to discover the risk factors of cardiovascular disease (CVD). The Offspring cohort of 5,124 participants was recruited in 1971. These individuals consist of the children and spouses of the children of the Original Cohort participants and attend the physical exams approximately every four years. The third generation (Gen3) Cohort, which consists of the grandchildren of the Original Cohort was enrolled in 2002. Till today, 32 exams have been performed in the Original Cohort, while 9 and 2 exams have been performed in the

Offspring and Gen3 Cohorts.

We perform a comparison using FHS GWAS and exome chip data with height as the outcome. Height was collected in exam 1 for all individuals in the three cohorts. FHS participants in the SNP Health Association Resource (SHARe) were genotyped using the Affymetrix 500K + 50K MIPS chip. We use these variants as the GWAS variants. As part of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, exome chip variants were genotyped with the Illumina HumanExome BeadChip [18]. The same filtering criteria are used to select variants for the PCs and GRMs computation:  $MAF \geq 5\%$ , call rate  $\geq 99\%$ , Hardy-Weinberg Equilibrium P-value  $> 0.0001$  and only one SNV of each pair of SNVs with LD  $r^2 > 0.5$  in a 50 SNVs window. Unrelated individuals are selected based on known pedigree structures to compute the weight of each SNV on PCs, then PCs are projected on related participants [24]. We test the association between the residual of height computed with adjustment for sex, age, age<sup>2</sup> and cohort indicator and SNV rs2322659, which is known for its spurious association with height due to population stratification. Two linear regression models,  $Y \sim SNV + PC_{GWAS}$  and  $Y \sim SNV + PC_{wes}$ , and two mixed effects models using GWAS-computed or exome chip-computed IBS kinship matrix, respectively, are performed with adjustment for the first 10 PCs. We then compare the effect size estimate and P-value for each model.

Table 2.1: Origin of 1000G EA and AA founders

<b>EA population</b>	<b>N</b>	<b>AA population</b>	<b>N</b>
FIN	99	ASW	61
CEU	95	ACB	96
GBR	90	LWK	97
IBS	107	ESN	99
TSI	107	GWD	113
		MSL	84
		YRI	107

### 2.3 Results

We select 498 unrelated EA individuals originating from 5 populations and 657 unrelated AA individuals originating from 7 populations in 1000G Phase 3 data set (Table 2.1). A total of 439,601 SNVs with  $MAF \geq 1\%$  for EA and AA, and 516,250 for EA only, pass the QC filters for type-I error evaluation. Among 439,601 SNVs in the EA and AA analysis, 180,472 SNVs in the GWAS set and 66,166 SNVs in the WES set pass the additional filtering criteria for PCs and GRM calculation. A total of 172,330 SNVs in the GWAS set and 57,584 SNVs in the WES set are included to compute PCs and GRM in the EA only analysis. A PC analysis restricted to unrelated 1000G EA and AA individuals, and EA individuals alone, is performed before we generate the simulated genotypes. Based on the clustering pattern in PCA, we divide the combined EA and AA, and EA samples alone, into 6 and 5 groups, respectively. Using the genotypes from 498 unrelated EA and 657 unrelated AA samples, 1,998 ( $333 \times 6$  groups) EA and 2,000 ( $400 \times 5$  groups) AA simulated samples are generated in each iteration of the simulation.

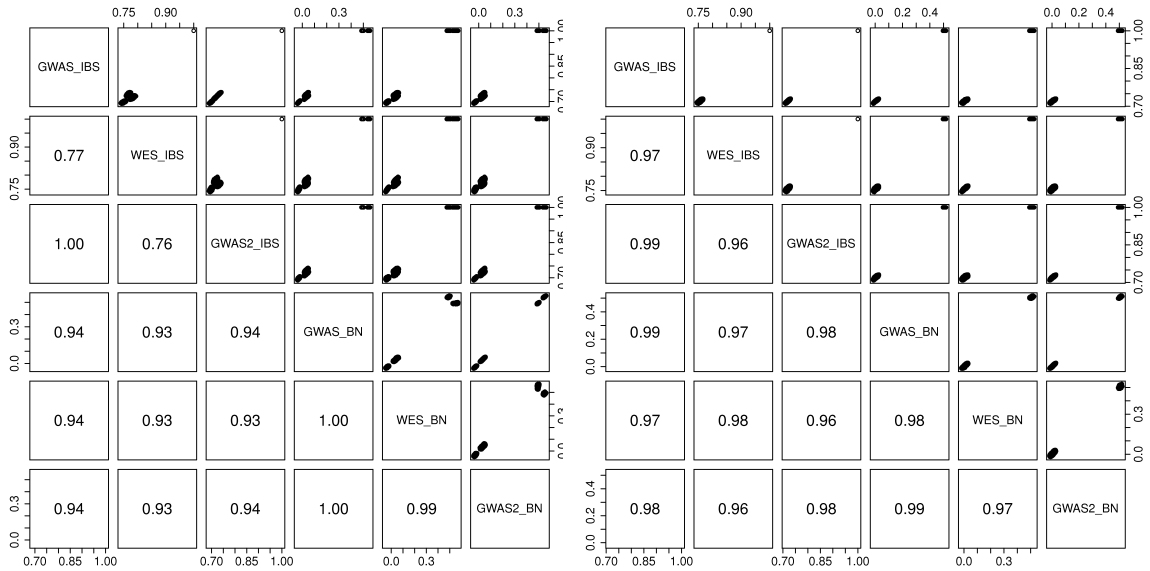
The Wilcoxon signed-rank test is performed on the Euclidean distance from the

group-specific centroid to each simulated sample. We make a pairwise comparison on distance difference computed using GWAS variants, WES variants and a randomly selected subset of GWAS variants which has the same number of variants as the WES set. Among 500 iterations, P-values of the Wilcoxon signed-rank test when comparing GWAS variants with WES variants and the randomly selected subset of GWAS variants are highly significant. P-values are  $2.9 \times 10^{-12}$  and  $1.6 \times 10^{-17}$  in the EA + AA dataset, and  $2.1 \times 10^{-4}$  and 0.01 in the EA only dataset, for the difference between GWAS and EC and GWAS and Random, respectively, which indicates that GWAS-computed PCs are significantly different from both WES-computed PCs and PCs computed using the subset of GWAS variants. However, only 100 iterations for EA + AA analysis and 26 for EA only analysis have P-value  $< 0.05$  when comparing WES-computed PCs with PCs computed using the subset of GWAS variants, and 35 iterations for EA + AA analysis and 5 for EA only analysis have P-value  $< 0.01$ . These results show that the differences between GWAS-computed and WES-computed PCs are mainly due to the number of variants included in the calculation.

In the comparison of kinship coefficients, Pearson correlations (Figure 2.2) show that kinship measures computed using GWAS and WES variants are highly correlated when there are only EA samples. In the presence of EA and AA samples, the correlation between WES and GWAS of the randomly selected set of GWAS markers is  $\sim 0.77$  in the

IBS kinship matrix, while it is above 0.9 in the BN kinship matrix.

Figure 2.2: Pair-wise comparison of kinship coefficients computed using GWAS and WES variants



Left: kinship coefficients in the combined EA and AA samples. Right: kinship coefficients in EA samples only. Plots above the diagonal are the scatterplots of the kinship coefficients. Plots below the diagonal are the Pearson correlations between them. GWAS\_IBS and GWAS\_BN represent the IBS and BN kinship matrix with GWAS markers, WES\_IBS and WES\_BN are the IBS and BN kinship matrix with WES markers, GWAS2\_IBS and GWAS2\_BN indicate the IBS and BN kinship matrix computed using a randomly selected subset of GWAS markers that has the same number of markers as the WES set

In both the EA + AA and EA only dataset, the genomic control factor  $\lambda_{GC}$  of unadjusted model  $Y \sim SNV$  indicates that there is population stratification present in the data, while in PC-adjusted models and LMM/GLMM-based methods,  $\lambda_{GC}$  falls within an acceptable range except in GLMMs using IBS kinship matrix alone (binary trait)(Table 2.2). This indicates that PCs and BN GRM computed using either GWAS or WES variants can

correct the population stratification in the data.

We use 439,601 SNVs (EA+AA) and 516,250 SNVs (EA only) in each iteration and repeat 500 times for a total of 219,800,500 (EA+AA) and 258,125,000 (EA only) unassociated SNVs in order to examine type-I error rate. For the continuous trait, we assign a pre-specified group-specific mean level, which is not associated with any simulated SNVs, plus a normally distributed random error term to each simulated individual. For the binary trait, we randomly select 999 cases and 999 controls (EA+AA), or 1000 cases and 1000 controls (EA only) among all individuals in simulation. In order to achieve the assumed population prevalence, we assume different prevalences across the subpopulations and select 34, 86, 138, 172, 259 and 310 cases out of 333 simulated samples in each of the 6 EA+AA group, and 50, 125, 200, 250, 375 cases out of 400 simulated samples in each of the 5 EA only group.

Type-I error rate is computed under 4 different significance levels: 0.05,  $10^{-3}$ ,  $10^{-4}$  and  $10^{-6}$  (Table 2.3). For the binary outcome, there is a deflation in type-I error rate in all models except the unadjusted model and GLMMs using the IBS kinship matrix alone. However, the two PC-adjusted models, two GLMMs using BN kinship matrix and four GLMMs using both GRM and PC adjustments have similar type-I error rate as the gold standard model, which shows that they can also correctly control the type-I error. For the continuous outcome, the PC-adjusted models correctly control type-I error. In LMMs, GRM computed using WES markers has higher type-I error rate than GRM computed

Table 2.2: Genomic control factor  $\lambda_{GC}$  from simulation studies

$\lambda_{GC}$	Trait	Model	EA+AA				EA only				
			Min	Median	Max	Min	Median	Max			
Binary		Y~SNV	43.66	43.98	44.24	6.052	6.117	6.225			
		Y~SNV+Group	0.987	1.003	1.021	0.986	1.003	1.019			
		Y~SNV+PCsGWAS	0.992	1.006	1.022	0.988	1.007	1.024			
		Y~SNV+PCsWES	0.991	1.007	1.024	0.992	1.007	1.025			
		Y~SNV+KinGWAS, IBS	1.065	1.087	1.091	1.089	1.105	1.122			
		Y~SNV+KinWES, IBS	1.040	1.063	1.087	1.066	1.087	1.108			
		Y~SNV+KinGWAS, BN	0.990	1.012	1.031	0.995	1.010	1.027			
		Y~SNV+KinWES, BN	0.982	1.000	1.020	0.983	1.007	1.023			
		Y~SNV+PCsGWAS+KinGWAS, IBS	0.986	1.004	1.013	0.987	1.001	1.019			
		Y~SNV+PCsWES+KinWES, IBS	0.989	1.003	1.014	0.989	1.001	1.013			
		Y~SNV+PCsGWAS+KinGWAS, BN	0.982	1.001	1.011	0.986	1.001	1.018			
		Y~SNV+PCsWES+KinWES, BN	0.994	1.002	1.019	0.988	1.000	1.010			
		Y~SNV	11.07	16.15	21.50	2.243	2.783	3.492			
	Cont.		Y~SNV+Group	0.986	1.000	1.021	0.981	0.999	1.018		
		Y~SNV+PCsGWAS	0.987	0.999	1.019	0.981	0.999	1.018			
		Y~SNV+PCsWES	0.985	1.000	1.023	0.982	0.999	1.017			
		Y~SNV+KinGWAS, IBS	0.995	1.009	1.027	1.002	1.018	1.035			
		Y~SNV+KinWES, IBS	0.992	1.003	1.018	0.987	1.005	1.026			
		Y~SNV+KinGWAS, BN	0.985	1.004	1.017	0.985	1.004	1.018			
		Y~SNV+KinWES, BN	0.990	1.008	1.029	0.995	1.013	1.035			

using GWAS markers. While the relative type-I error rates of both the GWAS- and WES-computed GRMs fall in an acceptable range at significance level 0.05,  $10^{-3}$  and  $10^{-4}$ , there is a small inflation in WES-computed GRM when the threshold is  $10^{-6}$ . The extent of inflation in the EA + AA analysis is less than that in the EA only analysis, which indicates that the GRM may not be sufficient to correct type-I error when the population stratification is more subtle. In either binary or continuous outcome, models including PCs computed using GWAS or WES variants do not show different type-I error rate.

Power is evaluated using 10 SNVs that are associated with the simulated continuous or binary phenotypes. In GLMMs for binary outcome, we focus on the models using the BN kinship matrix due to the inflated type-I error rate found in models with the IBS kinship matrix. We first compare power of using GWAS-computed PCs/GRM vs. WES-computed PCs/GRM. The empirical power evaluations are very similar in the PC- or LMM/GLMM-based methods between using GWAS-computed PCs/GRM and WES-computed PCs/GRM. Then we evaluate the performance of the PC- and LMM/GLMM-based methods between testing SNVs confounded and not confounded by population stratification. In PC-based models and GLMMs with PC adjustment, SNVs 1-5 (low weight, not confounded by population stratification) have higher power than SNVs 6-10 (high weight, confounded by population stratification) in general. This is exactly the result we expect because high weight SNVs contribute more to PCs than low weight SNVs and hence PCs can explain some of the phenotypic variance when testing high

Table 2.3: Relative type-I error rate in simulation studies

Relative type-I error rate	EA+AA					EA					
	$\alpha = 0.05$	$10^{-3}$	$10^{-4}$	$10^{-6}$	$\alpha = 0.05$	$10^{-3}$	$10^{-4}$	$10^{-6}$	$10^{-3}$	$10^{-4}$	$10^{-6}$
<b>Binary</b>											
Y~SNV	16.02	646.91	5724.46	$4.49 \times 10^5$	8.56	179.06	1101.56	$4.24 \times 10^4$	48.01	189.80	3025.58
Y~SNV+Group	0.99	0.92	0.89	0.87	0.99	0.93	0.87	0.71	1.00	1.00	0.99
Y~SNV+PCs <sub>GWAS</sub>	1.00	0.94	0.93	0.80	1.00	0.95	0.90	0.69	1.00	1.00	0.99
Y~SNV+PCs <sub>WES</sub>	1.01	0.96	0.94	0.83	1.00	0.95	0.91	0.71	1.00	1.00	0.99
Y~SNV+Kin <sub>GWAS</sub> , IBS	1.17	1.31	1.30	1.26	1.23	1.60	1.89	2.91	1.00	1.00	0.99
Y~SNV+Kin <sub>WES</sub> , IBS	1.11	1.18	1.15	1.07	1.19	1.46	1.66	2.40	1.00	1.00	0.99
Y~SNV+Kin <sub>GWAS</sub> , BN	1.00	0.91	0.80	0.53	1.01	0.97	0.94	0.83	1.00	1.00	0.99
Y~SNV+Kin <sub>WES</sub> , BN	0.98	0.86	0.76	0.56	1.00	0.95	0.91	0.71	1.00	1.00	0.99
Y~SNV+PCs <sub>GWAS</sub> +Kin <sub>GWAS</sub> , IBS	1.00	1.00	0.98	0.78	1.00	0.96	0.91	0.66	1.00	1.00	0.99
Y~SNV+PCs <sub>WES</sub> +Kin <sub>WES</sub> , IBS	1.00	1.01	0.99	0.92	1.00	0.95	0.90	0.70	1.00	1.00	0.99
Y~SNV+PCs <sub>GWAS</sub> +Kin <sub>GWAS</sub> , BN	1.00	0.98	0.97	0.87	1.00	0.96	0.90	0.67	1.00	1.00	0.99
Y~SNV+PCs <sub>WES</sub> +Kin <sub>WES</sub> , BN	1.00	1.00	0.96	0.98	1.00	0.95	0.90	0.64	1.00	1.00	0.99
<b>Continuous</b>											
Y~SNV	12.75	382.99	2839.61	$1.47 \times 10^5$	4.87	48.01	189.80	3025.58	48.01	189.80	3025.58
Y~SNV+Group	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99
Y~SNV+PCs <sub>GWAS</sub>	1.00	1.00	1.01	1.02	1.00	1.00	1.00	0.99	1.00	1.00	0.99
Y~SNV+PCs <sub>WES</sub>	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.07	1.00	1.00	0.99
Y~SNV+Kin <sub>GWAS</sub> , IBS	1.01	1.03	1.03	1.09	1.01	1.04	1.06	1.14	1.00	1.00	0.99
Y~SNV+Kin <sub>WES</sub> , IBS	1.02	1.06	1.08	1.16	1.04	1.11	1.16	1.38	1.00	1.00	0.99
Y~SNV+Kin <sub>GWAS</sub> , BN	1.01	1.03	1.03	1.04	1.00	1.02	1.05	1.17	1.00	1.00	0.99
Y~SNV+Kin <sub>WES</sub> , BN	1.02	1.05	1.07	1.15	1.03	1.09	1.14	1.28	1.00	1.00	0.99

A ratio of observed type-I error rate to expected type-I error rate for various P-value thresholds are presented. A ratio  $> 1$  shows inflation and a ratio  $< 1$  shows deflation.

weight SNVs, which in turn decreases power in association tests. In GLMMs using GRM alone (binary outcome), high weight SNVs achieve higher power than low weight SNVs. In LMMs (continuous outcome), high weight SNVs also have higher power when there are EA samples only, while low weight SNVs have higher power in the simulation of EA and AA samples. Finally, we directly compare the performance of PC-based vs. LMM/GLMM-based methods. For SNVs 6-10, LMMs and GLMMs using GRM alone outperform PC-based models in either the analysis with EA and AA individuals or the analysis restricted to EA individuals. For SNVs 1-5, PC-based models and LMMs have similar power in general when the phenotype is continuous. They also have comparable performance in the analysis including EA and AA samples when the phenotype is binary and a slightly higher power is achieved in PC-based models with EA samples only. In addition, GLMMs using both GRM and PC adjustment have similar power to the PC-based models for binary outcome.

In the application of FHS data, a subset of 122,233 SNVs in the GWAS set and 18,107 SNVs in the exome chip set pass the filtering criteria. A total of 2,464 unrelated individuals are selected to compute PC weights based on the known pedigree structure. The association analyses include 7,269 individuals in total. P-values in the unadjusted model and LMM with a kinship matrix computed using the pedigree are  $1.6 \times 10^{-15}$  and 0.03 respectively, which indicates SNV rs2322659 is strongly associated with height. However this association disappears in PC-adjusted models and LMMs using an empirical

Table 2.4: Power from simulation studies in EA + AA samples

Power (%)	Low loading										High loading											
	Model	SNV1	SNV2	SNV3	SNV4	SNV5	SNV6	SNV7	SNV8	SNV9	SNV10	SNV1	SNV2	SNV3	SNV4	SNV5	SNV6	SNV7	SNV8	SNV9	SNV10	
	<b>Binary</b>																					
	$Y \sim \text{SNV}$	53.2	83.1	73.9	72.3	53.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	$Y \sim \text{SNV} + \text{Group}$	46.1	60.2	52.2	52.1	52.6	41.1	31.9	28.9	36.1	47.2											
	$Y \sim \text{SNV} + \text{PC}_{\text{S}}^{\text{GWAS}}$	45.5	60.2	52.7	53.1	51.9	41.1	27.6	29.9	34.3	46.2											
	$Y \sim \text{SNV} + \text{PC}_{\text{S}}^{\text{WES}}$	45.6	61.1	53.2	54.8	51.8	40.2	31.3	28.0	35.5	47.2											
	$Y \sim \text{SNV} + \text{Kin}_{\text{GWAS}}, \text{IBS}$	49.6	50.4	51.7	44.4	49.4	57.9	65.6	57.9	59.8	57.7											
	$Y \sim \text{SNV} + \text{Kin}_{\text{WES}}, \text{IBS}$	50.8	52.8	55.8	45.5	52.8	52.5	58.1	51.8	56.1	50.8											
	$Y \sim \text{SNV} + \text{Kin}_{\text{GWAS}}, \text{BN}$	49.0	51.0	51.8	45.4	49.2	54.4	58.1	50.6	53.4	53.4											
	$Y \sim \text{SNV} + \text{Kin}_{\text{WES}}, \text{BN}$	50.5	52.4	54.9	44.8	52.1	48.3	52.1	47.6	52.1	45.4											
	$Y \sim \text{SNV} + \text{PC}_{\text{S}}^{\text{GWAS}} + \text{Kin}_{\text{GWAS}}, \text{IBS}$	52.1	53.3	53.9	53.0	54.9	37.2	26.8	23.3	41.0	34.9											
	$Y \sim \text{SNV} + \text{PC}_{\text{S}}^{\text{WES}} + \text{Kin}_{\text{WES}}, \text{IBS}$	53.6	52.4	53.6	52.1	54.3	38.2	28.8	23.6	39.0	35.6											
	$Y \sim \text{SNV} + \text{PC}_{\text{S}}^{\text{GWAS}} + \text{Kin}_{\text{GWAS}}, \text{BN}$	51.1	51.3	51.1	50.6	52.1	40.4	29.5	25.9	39.3	36.8											
	$Y \sim \text{SNV} + \text{PC}_{\text{S}}^{\text{WES}} + \text{Kin}_{\text{WES}}, \text{BN}$	51.0	52.6	54.0	52.0	54.0	37.4	27.5	24.2	40.4	34.8											
	<b>Continuous</b>																					
	$Y \sim \text{SNV}$	63.5	66.9	74.8	73.0	57.5	99.7	100.0	99.7	100.0	99.7											
	$Y \sim \text{SNV} + \text{Group}$	60.6	61.9	66.7	64.6	61.4	53.3	53.0	50.7	53.3	50.9											
	$Y \sim \text{SNV} + \text{PC}_{\text{S}}^{\text{GWAS}}$	61.9	58.8	65.4	63.5	61.2	53.3	52.8	52.8	52.8	50.9											
	$Y \sim \text{SNV} + \text{PC}_{\text{S}}^{\text{WES}}$	60.6	60.6	66.4	64.3	61.4	53.0	53.3	52.2	53.5	50.7											
	$Y \sim \text{SNV} + \text{Kin}_{\text{GWAS}}, \text{IBS}$	61.7	61.7	68.0	64.3	61.2	57.0	59.3	57.2	59.6	57.0											
	$Y \sim \text{SNV} + \text{Kin}_{\text{WES}}, \text{IBS}$	62.2	61.7	68.8	65.4	60.4	57.7	57.0	54.6	58.0	57.5											
	$Y \sim \text{SNV} + \text{Kin}_{\text{GWAS}}, \text{BN}$	61.7	60.9	68.5	64.3	60.4	57.5	59.8	56.4	60.1	57.0											
	$Y \sim \text{SNV} + \text{Kin}_{\text{WES}}, \text{BN}$	62.2	61.9	69.3	65.1	59.8	57.7	57.0	54.3	58.0	57.5											

Table 2.5: Power from simulation studies in EA samples

Power (%)	Low loading										High loading											
	Model	SNV1	SNV2	SNV3	SNV4	SNV5	SNV6	SNV7	SNV8	SNV9	SNV10	SNV1	SNV2	SNV3	SNV4	SNV5	SNV6	SNV7	SNV8	SNV9	SNV10	
	<b>Binary</b>																					
	Y~SNV	51.5	56.0	52.2	64.1	62.0	82.5	99.0	100.0	100.0	100.0	99.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
	Y~SNV+Group	54.4	56.5	57.1	58.3	57.4	42.5	58.8	51.0	41.5	47.2	58.8	51.0	41.5	47.2	42.5	58.8	51.0	41.5	47.2	47.2	
	Y~SNV+PCs <sub>GWAS</sub>	52.0	56.2	55.8	58.4	59.5	45.3	55.9	51.2	41.7	47.0	55.9	51.2	41.7	47.0	45.3	55.9	51.2	41.7	47.0	47.0	
	Y~SNV+PCs <sub>WES</sub>	55.8	55.1	57.5	58.9	58.2	43.5	59.2	53.1	33.4	46.3	59.2	53.1	33.4	46.3	43.5	59.2	53.1	33.4	46.3	46.3	
	Y~SNV+King <sub>GWAS</sub> , IBS	47.0	47.5	41.9	48.5	56.4	51.7	90.0	86.2	89.0	90.7	90.0	86.2	89.0	90.7	51.7	90.0	86.2	89.0	90.7	90.7	
	Y~SNV+Kin <sub>WES</sub> , IBS	47.1	47.1	43.1	48.9	55.0	51.3	90.2	86.4	86.7	89.2	90.2	86.4	86.7	89.2	51.3	90.2	86.4	86.7	89.2	89.2	
	Y~SNV+King <sub>GWAS</sub> , BN	46.4	45.2	43.0	48.8	56.4	49.0	84.2	82.8	85.6	87.6	84.2	82.8	85.6	87.6	49.0	84.2	82.8	85.6	87.6	87.6	
	Y~SNV+Kin <sub>WES</sub> , BN	46.8	45.7	42.3	48.3	56.4	47.9	83.3	83.0	84.3	86.7	83.3	83.0	84.3	86.7	47.9	83.3	83.0	84.3	86.7	86.7	
	Y~SNV+PCs <sub>GWAS</sub> +Kin <sub>GWAS</sub> , IBS	57.9	56.0	56.2	54.1	67.4	46.6	48.9	53.6	39.9	46.6	48.9	53.6	39.9	46.6	67.4	46.6	48.9	53.6	39.9	46.6	
	Y~SNV+PCs <sub>WES</sub> +Kin <sub>WES</sub> , IBS	57.2	56.3	56.8	54.9	66.0	45.8	48.7	54.9	31.8	49.2	48.7	54.9	31.8	49.2	66.0	45.8	48.7	54.9	31.8	49.2	
	Y~SNV+PCs <sub>GWAS</sub> +Kin <sub>GWAS</sub> , BN	57.8	55.7	56.1	54.0	67.6	46.8	48.5	53.6	40.7	46.4	48.5	53.6	40.7	46.4	67.6	46.8	48.5	53.6	40.7	46.4	
	Y~SNV+PCs <sub>WES</sub> +Kin <sub>WES</sub> , BN	57.6	56.7	56.2	55.3	66.3	46.1	48.9	54.1	31.4	48.5	48.9	54.1	31.4	48.5	66.3	46.1	48.9	54.1	31.4	48.5	
	<b>Continuous</b>																					
	Y~SNV	61.6	60.6	61.6	62.4	62.4	74.9	91.4	92.5	91.4	95.3	91.4	92.5	91.4	95.3	62.4	74.9	91.4	92.5	91.4	95.3	
	Y~SNV+Group	64.5	60.6	64.2	60.9	67.4	59.9	60.2	62.0	55.6	61.6	60.2	62.0	55.6	61.6	67.4	59.9	60.2	62.0	55.6	61.6	
	Y~SNV+PCs <sub>GWAS</sub>	64.5	62.0	64.5	62.0	66.3	60.2	59.9	60.9	54.1	60.9	60.2	60.9	54.1	60.9	66.3	60.2	59.9	60.9	54.1	60.9	
	Y~SNV+PCs <sub>WES</sub>	64.5	61.6	65.6	61.6	67.7	60.9	60.2	61.6	48.4	59.5	60.2	61.6	48.4	59.5	67.7	60.9	60.2	61.6	48.4	59.5	
	Y~SNV+King <sub>GWAS</sub> , IBS	62.7	60.6	63.1	60.6	66.7	62.0	71.7	72.4	68.1	72.0	71.7	72.4	68.1	72.0	66.7	62.0	71.7	72.4	68.1	72.0	
	Y~SNV+Kin <sub>WES</sub> , IBS	64.2	61.3	62.7	60.2	66.3	62.0	69.9	69.5	66.3	71.7	69.9	69.5	66.3	71.7	66.3	62.0	69.9	69.5	66.3	71.7	
	Y~SNV+King <sub>GWAS</sub> , BN	62.7	60.6	63.1	60.6	66.7	60.9	71.0	71.3	65.2	73.1	71.0	71.3	65.2	73.1	66.7	60.9	71.0	71.3	65.2	73.1	
	Y~SNV+Kin <sub>WES</sub> , BN	64.2	61.3	63.4	60.2	66.3	61.6	69.2	69.5	65.6	70.6	69.2	69.5	65.6	70.6	66.3	61.6	69.2	69.5	65.6	70.6	

Table 2.6: Association test results from FHS

<b>Model</b>	$\beta$	<b>SE</b>	<b>P-value</b>
Height $\sim$ SNV	-0.1343	0.017	$1.6 \times 10^{-15}$
Height $\sim$ SNV+Kin <sub>pedigree</sub>	0.0893	0.042	0.03
Height $\sim$ SNV+Kin <sub>pedigree</sub> +PC <sub>SGWAS</sub>	0.0031	0.046	0.95
Height $\sim$ SNV+Kin <sub>pedigree</sub> +PC <sub>SEC</sub>	-0.0019	0.046	0.97
Height $\sim$ SNV+Kin <sub>GWAS</sub>	0.0220	0.020	0.27
Height $\sim$ SNV+Kin <sub>EC</sub>	0.0248	0.019	0.19

kinship matrix (Table 2.6). This confirms that the spurious association in the unadjusted model is due to population stratification. P-values of the PC-adjusted models and LMMs using an empirical kinship matrix are all above 0.05, which shows that the adjustment using exome chip variants can also alleviate the spurious association due to population stratification in practice.

## 2.4 Discussion

This paper sought to compare PCs and GRM computed using GWAS and WES markers, and to evaluate model performance of PC-based and LMM/GLMM-based methods in terms of type-I error rate and power in association tests. Intuitively, in studies with WES data, adjustment for potential population stratification may not be achieved because WES markers only contain ancestry information within the exome. WES-computed PCs and GRM may not be able to capture all useful information available in whole genome data. Besides this concern, the fewer number of WES variants may also lead to insufficient adjustment of population stratification because the best ancestry estimates are usually obtained using a very large number of random markers [24]. However, our simulation and

real data example showed that WES markers can achieve a similar performance as GWAS markers for population stratification adjustment.

Through the comparison between PCs computed from GWAS and WES markers, we found that the significant P-value in Wilcoxon signed rank test was due to the difference in the number of GWAS and WES variants included in PCA. When we used the same number of GWAS and WES variants to compute PCs, the difference disappeared. Comparison among kinship demonstrated that the BN GRM showed consistent high correlation ( $\sim 0.99$ ) between GWAS and WES markers, while GWAS- and WES-computed IBS GRMs were less correlated ( $\sim 0.76$ ). It suggests that the WES GRM may not perform as well as the GWAS GRM in association testing, which was verified in our type-I error and power simulations.

Genomic control factor is often used to examine the inflation in the middle of the null distribution, while type-I error rate is used to examine the tail of it. We compared both quantities through simulation and drew different conclusions about the ancestry adjustment using WES variants. The genomic control factor showed no evidence of population stratification in models using either GWAS- or WES-computed PCs or GRM (except IBS GRM in binary trait). However, a slightly inflated type-I error rate was found in LMMs (continuous trait) with WES-computed GRM at a  $10^{-6}$  significance level. In contrast to the WES-computed GRM, the GWAS-computed GRM showed no inflated

type-I error rate and had similar performance as the PC-adjusted models in these two quantities. These results reflect the medium level of correlation ( $\sim 0.76$ ) between GWAS- and WES-computed kinship coefficients.

Our power evaluation was conducted using 5 SNVs that were confounded by population stratification and 5 SNVs that were not confounded by stratification. The power of using WES-computed PCs/GRM was very close to GWAS-computed PCs/GRM, which indicates that it is appropriate to use WES data to detect SNVs with true effect.

PC-adjusted models had higher power for SNVs not confounded by population stratification and lower power for SNVs confounded by population stratification. High weight SNVs contributed more to PCs than low weight SNVs and hence PCs can explain part of the phenotypic variance when testing high weight SNVs, which in turn decreases the power in association analyses. A similar result was also found in LMMs in the presence of EA and AA individuals, which is consistent with the previous findings that the candidate marker should not be included in the GRM due to a potential loss of power [25]. LMMs had similar power as PC-based models in tests of low weight SNVs and higher power when testing high weight SNVs. Hence, LMMs are more appropriate in association analyses due to their better performance on high weight SNVs when the phenotype is continuous. For the binary outcome, GLMMs performed better when testing high weight

SNVs.

Another interesting finding is that the number of variants included in PCA determines the size of the clustering. When more SNVs are included in PC computation, the Euclidean distance between PCs of two individuals sharing the same ancestry becomes smaller, while the distance between two samples from different ancestry becomes larger.

Although using GWAS variants to compute PCs should be the preferred approach because it captures variation over the whole genome, WES-computed PCs are sufficient to control inflated type-I error due to population stratification and provide similar power to approaches adjusting for GWAS-computed PCs. For the continuous phenotype, LMMs should be preferred over PC-adjusted models because they perform better if associated SNVs are confounded with population stratification, such as SNVs in the HLA regions, which are associated with many auto-immune traits but also show signs of population stratification.

## Chapter 3 A Family-Informed Phenotype Imputation Approach

### 3.1 Introduction

GWAS have been very successful in detecting SNVs associated with complex traits. The power of GWAS is limited by the number of individuals with data available for the trait of interest. For easily measured traits, tens of thousands of individuals are typically contributing to GWAS. However, a lack of statistical power can still occur because of missingness in phenotypic data. Some phenotypes are difficult to collect due to cost, loss to follow-up and inaccessibility of the biological sample at the time of the study. Removing the samples with missing data will decrease sample size and may introduce bias. However, with the increased use of joint analysis of multiple phenotypes and Phenome-wide association studies (PheWAS) [36], it has become possible to collect a large dataset on many correlated phenotypes so that the effect of missing data on one particular phenotype can be minimized by utilizing additional correlated phenotypes. In addition, for studies with family data, relatives are also a good source of information for missing values on heritable traits and should be leveraged in the phenotype imputation process.

Multivariate normal (MVN) distribution has been widely used in modeling the distribution of the observed and missing data. A conditional MVN distribution (imputed values conditioning on observed values) can then be obtained to generate imputed values. Price et al. [35] imputed the Z statistic of untyped marker in an association test by exploiting the Z statistic of typed markers and the linkage disequilibrium (LD) pattern between them

through a conditional MVN distribution. They showed that their method can recover 84% of the effective sample size for common variants ( $MAF > 5\%$ ) and 54% for rare variants ( $1\% < MAF < 5\%$ ). PhenIMP [37] is a phenotype imputation approach developed for GWAS with unrelated samples. PhenIMP uses the correlation between phenotypes to impute the missing phenotype values which are assumed to follow a conditional MVN distribution given observed values. The pair-wise correlation matrix between phenotypes is estimated from a pilot dataset prior to the analysis. Yet another Bayesian approach PHENIX [38] also assumes that the approximate posterior distribution of the missing values has the form of a MVN distribution. PHENIX treats all missing phenotype values as parameters in a Bayesian mixed model to derive the posterior distribution of the parameters using a Variational Bayes (VB) approach.

We propose a new family-informed phenotype imputation method which uses the polygenic and environmental variance components of the phenotypes and the family structure of the samples. Missing values are imputed from three pieces of information: correlated phenotype values of the missing individual, missing phenotype values of the missing individual's relatives and the correlated phenotype values of these relatives. Our method can be applied not only in studies with family data, but also in population-based studies by using an empirical kinship matrix to account for cryptic relatedness.

In this paper, we first propose a family-informed phenotype imputation approach which

utilizes family structure and additional correlated phenotypes. We show that the imputation accuracy can be improved by including family structure in the imputation. We then explore several features of PhenIMP for studies with unrelated samples. We used extensive simulations to evaluate the performance of our family-informed method and verify our conclusions about PhenIMP. Specifically, we identify the situations where imputation can boost power in an association test and situations where imputation does not result in a gain in power, which should be very helpful when designing a study.

## 3.2 Methods

### 3.2.1 Phenotype Imputation for Family Data

Assume  $K$  phenotypes are collected on  $n$  individuals and missingness can occur in any of the  $K$  phenotypes. Let  $Y_k$  be a vector of length  $n$  to represent values of the  $k$ th phenotype and  $\mathbf{Y}=(Y_1, \dots, Y_K)^T$ . We assume that the phenotype vector  $\mathbf{Y}$  follows a MVN distribution. The expectation vector of  $\mathbf{Y}$  is  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^T$ , where  $\mu_k$  is the vector with each element being the mean of phenotype  $k$ . We partition the unconditional covariance of  $\mathbf{Y}$  into the polygenic variance component and the environmental variance component as  $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_A \otimes \boldsymbol{\Phi} + \boldsymbol{\Sigma}_E \otimes \mathbf{I}$ , where  $\otimes$  represents the Kronecker product of two matrices and

$$\boldsymbol{\Sigma}_A = \begin{pmatrix} \sigma_{A11}^2 & \sigma_{A12} & \cdots & \sigma_{A1K} \\ \sigma_{A12} & \sigma_{A22}^2 & \cdots & \sigma_{A2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{AK1} & \sigma_{AK2} & \cdots & \sigma_{AKK}^2 \end{pmatrix}, \boldsymbol{\Sigma}_E = \begin{pmatrix} \sigma_{E11}^2 & \sigma_{E12} & \cdots & \sigma_{E1K} \\ \sigma_{E12} & \sigma_{E22}^2 & \cdots & \sigma_{E2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{EK1} & \sigma_{EK2} & \cdots & \sigma_{EKK}^2 \end{pmatrix}, \quad (3.1)$$

Let  $\sigma_{A_{kk}}^2$  and  $\sigma_{E_{kk}}^2$  indicate the polygenic and environmental variance of phenotype  $k$ , respectively, and  $\sigma_{A_{kl}}$  and  $\sigma_{E_{kl}}$  indicate the polygenic and environmental covariance between phenotypes  $k$  and  $l$ , respectively. All of these quantities can be estimated using the maximum likelihood estimator (MLE) as implemented in the SOLAR software [23]. The kinship matrix  $\Phi$  describes the coefficient of relationship between two individuals. It can be derived using pedigree structure in family studies or empirically estimated using genotypes. The distribution of  $\mathbf{Y}$  is represented compactly using a block matrix

$$\begin{pmatrix} Y_m \\ Y_o \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_m \\ \mu_o \end{pmatrix}, \begin{pmatrix} \Sigma_{mm} & \Sigma_{mo} \\ \Sigma_{om} & \Sigma_{oo} \end{pmatrix}\right) \quad (3.2)$$

where,  $Y_m$  is the vector of all missing values in the phenotype vector  $Y$  (missing data) and  $Y_o$  is the vector of all remaining elements (observed data). The parameters  $\mu_m$  and  $\mu_o$  are the corresponding vectors of the expectation  $\mu$  and  $\Sigma_{mm}$ ,  $\Sigma_{mo}$ ,  $\Sigma_{om}$  and  $\Sigma_{oo}$  are the corresponding block matrices of  $\Sigma$ .

The conditional distribution of  $Y_m|Y_o$  follows a MVN distribution, where the mean is computed as

$$E(Y_m|Y_o) = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}(Y_o - \mu_o) \quad (3.3)$$

To estimate  $E(Y_m|Y_o)$ , we use the sample mean of the observed data to estimate  $\mu_m$  and  $\mu_o$ , and the MLE of the polygenic and environmental variance of the phenotypes along with the kinship matrix to estimate  $\Sigma_{mo}$  and  $\Sigma_{oo}$ . We then use the estimated  $E(Y_m|Y_o)$  as

the imputed values for the missing data in all phenotypes. The second term in the above equation shows that the imputed value depends on the family structure of the missing sample, expressed by  $\Sigma_{mo}\Sigma_{oo}^{-1}$ , and the observed phenotypes of the missing sample and the relatives by  $Y_o - \mu_o$ .

### 3.2.2 Phenotype Imputation for Population-Based Studies

In a population-based study, we can assume that the individuals are unrelated, that is  $\Phi = I$ . The above polygenic model can therefore be simplified, which is equivalent to the approach implemented in PhenIMP. We use  $y_{k,i}$  to represent the  $i$ th element in vector  $Y_k$ . The simplified model can be written as

$$\begin{pmatrix} y_{1,i} \\ \vdots \\ y_{K,i} \end{pmatrix} \sim MVN\left( \begin{pmatrix} \mu_{1,i} \\ \vdots \\ \mu_{K,i} \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{12} & \sigma_{22}^2 & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_{KK}^2 \end{pmatrix} \right), \quad (3.4)$$

where  $\mu_{k,i}$  is the  $i$ th element of  $\mu_k$  (mean of phenotype  $k$ ),  $\sigma_{kk}^2$  is the variance of phenotype  $k$  and  $\sigma_{kl}$  is the covariance between phenotype  $k$  and  $l$ . Similarly, we can move all the missing values of individual  $i$  to  $Y_m$  and all the observed values to  $Y_o$  and rewrite the phenotype vector as

$$\begin{pmatrix} Y_m \\ Y_o \end{pmatrix} \sim MVN\left( \begin{pmatrix} \mu_m \\ \mu_o \end{pmatrix}, \begin{pmatrix} \Sigma_{mm} & \Sigma_{mo} \\ \Sigma_{om} & \Sigma_{oo} \end{pmatrix} \right) \quad (3.5)$$

and the imputed values are estimated  $E(Y_m|Y_o) = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}(Y_o - \mu_o)$ .

### 3.2.3 Relationship with Imputation Using Regression Model

For simplicity, consider the samples are unrelated and missingness only occurs in phenotype 1, that is,  $Y_1$  is observed for the first  $n(1-r)$  elements and missing for the last  $nr$  elements, where  $r$  is the percentage of missing data, and  $Y_2, \dots, Y_K$  are complete. In imputation using a regression model, a regression model with  $Y_1$  as the dependent variable and  $Y_2, \dots, Y_K$  as the independent variable is fitted using the  $n(1-r)$  samples with complete information, then the  $nr$  missing values on phenotype 1 are imputed using predictions from the estimated regression model by plugging in the corresponding elements from  $Y_2, \dots, Y_K$ .

Therefore, from conditional mean imputation, the estimated regression coefficients (without intercept) are

$$\hat{\beta} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Y} \quad (3.6)$$

where  $\mathcal{X}_{n(1-r) \times (K-1)} = \begin{pmatrix} y_{2,1} - \bar{y}_2 & y_{3,1} - \bar{y}_3 & \cdots & y_{K,1} - \bar{y}_K \\ y_{2,2} - \bar{y}_2 & y_{3,2} - \bar{y}_3 & \cdots & y_{K,2} - \bar{y}_K \\ \vdots & \ddots & & \vdots \\ y_{2,n(1-r)} - \bar{y}_2 & y_{3,n(1-r)} - \bar{y}_3 & \cdots & y_{K,n(1-r)} - \bar{y}_K \end{pmatrix}$  and

$$\mathcal{Y}_{n(1-r) \times 1} = \begin{pmatrix} y_{1,1} - \bar{y}_1 \\ y_{1,2} - \bar{y}_1 \\ \vdots \\ y_{1,n(1-r)} - \bar{y}_1 \end{pmatrix}.$$

For  $i = n(1-r) + 1, \dots, n$ , the imputed values of phenotype 1 from imputation using a regression model are

$$y_{1,i} = \bar{y}_1 + \hat{\beta}^T (y_{2,i} - \bar{y}_2, y_{3,i} - \bar{y}_3, \dots, y_{K,i} - \bar{y}_K)^T \quad (3.7)$$

The imputed values from PhenIMP are

$$y_{1,i} = \bar{y}_1 + \Sigma_{mo} \Sigma_{oo}^{-1} (y_{2,i} - \bar{y}_2, y_{3,i} - \bar{y}_3, \dots, y_{K,i} - \bar{y}_K)^T \quad (3.8)$$

where  $\Sigma_{mo} = (\text{cov}(y_1, y_2), \text{cov}(y_1, y_3), \dots, \text{cov}(y_1, y_K))$  and

$$\Sigma_{oo} = \begin{pmatrix} \text{cov}(y_2, y_2) & \text{cov}(y_2, y_3) & \cdots & \text{cov}(y_2, y_K) \\ \text{cov}(y_3, y_2) & \text{cov}(y_3, y_3) & \cdots & \text{cov}(y_3, y_K) \\ \vdots & \ddots & & \vdots \\ \text{cov}(y_K, y_2) & \text{cov}(y_K, y_3) & \cdots & \text{cov}(y_K, y_K) \end{pmatrix}$$

Note that  $n\Sigma_{oo} = \mathcal{X}^T \mathcal{X}$  and  $n\Sigma_{mo} = \mathcal{Y}^T \mathcal{X}$  when  $n$  is large. We have

$\hat{\beta}^T = \mathcal{Y}^T \mathcal{X} (\mathcal{X}^T \mathcal{X})^{-1} = \Sigma_{mo} \Sigma_{oo}^{-1}$ , hence PhenIMP is equivalent to imputation using a

regression model.

### 3.2.4 Power of Single Variant Test

To compute the theoretical statistical power for a population-based GWAS, we derive the non-centrality parameter (NCP) for the test statistic in a single variant association test (linear regression) using the imputed phenotype, without assuming the noisy measurement model (NMM), which was introduced in Hormozdiari et al. [37]. Without loss of generality, we assume that all covariates are already adjusted using a linear regression model and the residuals are used in the imputation and association test. We also center the residuals so they have mean 0.

In NMM, phenotype 1 is assumed to have the strongest association with the SNV tested. Other phenotypes driven by a smaller genetic effect can then be modeled as phenotype 1 plus noise. We assume phenotype  $k$  is one of these phenotypes. When testing the effect of the SNV on phenotype 1, the test statistic can be written as  $s_1 \sim N(\frac{\beta_1}{\sigma_1} \sqrt{n}, 1)$ , as introduced in Chapter 1. Under NMM, the test statistic of phenotype  $k$  can then be written as  $s_k \sim N(r_{1k} \frac{\beta_1}{\sigma_1} \sqrt{k}, 1)$ , where  $r_{1k}$  is the correlation between phenotypes 1 and  $k$ . In phenotype imputation, the imputed phenotype values can be considered as the unobserved true phenotype values plus noise. Hence, the NCP of the test statistic for the imputed phenotype can be approximated by the product of the NCP of the test statistic for the

unobserved true phenotype and the correlation between the imputed and true phenotypes.

NCP is the expectation of the test statistic, hence when we perform association testing on all  $nr$  imputed individuals, we have

$$NCP = E\left(\frac{\hat{\beta}}{SE(\hat{\beta})}\right) = \frac{\beta}{\sqrt{\sigma^2/2p(1-p)nr}} \quad (3.9)$$

where  $p$  is the coded allele frequency of the SNV tested and  $\sigma^2$  is the variance of phenotype 1. Some simple math derivation can show that  $\hat{\beta} = \Sigma_{mo}\Sigma_{oo}^{-1}(\hat{\beta}_2, \dots, \hat{\beta}_K)^T$  and  $\sigma^2 = \Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}$ , where  $\hat{\beta}_k$  represents the estimated effect size of the  $k$ th phenotype.

Therefore, for the association test of  $nr$  imputed samples, we have

$$NCP = \frac{\Sigma_{mo}\Sigma_{oo}^{-1}(\beta_2, \dots, \beta_K)^T}{\sqrt{\Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}}} \sqrt{2p(1-p)nr} \quad (3.10)$$

where  $\beta_k$  is the true effect size of the SNV on phenotype  $k$ .

### 3.2.5 Analysis with Combined Observed and Imputed Data

Besides performing an association test on imputed data only, it might be desirable to analyze the observed data along with the imputed data to maximize the sample size and hence improve power of the test. Hormozdiari et al. [37] proposed to use Stouffer's signed Z-score method to meta-analyze the observed and imputed data, with an optimal weight to

account for the imputation accuracy. However, the Stouffer's signed Z-score method does not provide a pooled effect size estimate. Because the individual level data are available, it is possible to perform an association test on the pooled data set. We compare these two ways of combining the observed and imputed data, and compute the NCP to approximate statistic power theoretically.

In Stouffer's signed Z-score method, the test statistic is  $s_{meta} = \frac{w_{obs}s_{obs} + w_{imp}s_{imp}}{\sqrt{w_{obs}^2 + w_{imp}^2}}$ , where  $s_{obs}$  and  $s_{imp}$  are the test statistics from the observed and imputed data, respectively,

$w_{obs} = \sqrt{n(1-r)}$  and  $w_{imp} = \sqrt{nr}$  are the weights for each test statistic, which are the square root of the sample sizes. Hormozdiari et al. [37] proposed an optimal weight to achieve the maximum NCP:  $w_{obs} = \sqrt{n(1-r)}$  and  $w_{imp} = \sqrt{\Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}nr}$ . Therefore, the NCP of Stouffer's signed Z-score method is

$$NCP_{meta} = \frac{(1-r)\beta_1/\sigma_1 + r\Sigma_{mo}\Sigma_{oo}^{-1}(\beta_2, \dots, \beta_K)^T}{\sqrt{(1-r) + r\Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}}} \sqrt{2p(1-p)n} \quad (3.11)$$

Next, to compute the test statistic of the pooled data, we can show that

$E(\hat{\beta}) = (1-r)\beta_1 + r\Sigma_{mo}\Sigma_{oo}^{-1}(\beta_2, \dots, \beta_K)^T$  and  $\sigma^2 = (1-r)\sigma_1^2 + r\Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}$ , where  $\beta_k$  is the effect size of the SNV on the  $k$ th phenotype and  $\sigma_1^2$  is the variance of the first phenotype on all observed samples. Hence, the NCP of the single variant test using both

the combined observed and imputed samples is computed as

$$NCP_{pooled} = \frac{(1-r)\beta_1 + r\Sigma_{mo}\Sigma_{oo}^{-1}(\beta_2, \dots, \beta_K)^T}{\sqrt{(1-r)\sigma_1^2 + r\Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}}} \sqrt{2p(1-p)n} \quad (3.12)$$

Comparing  $NCP_{meta}$  with  $NCP_{pooled}$ , we can see that these two approaches are identical when  $\sigma_1^2 = 1$ , which means the data are standardized.

### 3.2.6 Strategy to Analyze the Observed and Imputed Data

In the imputed dataset, the expected value of the effect size in the pooled analysis is  $\Sigma_{mo}\Sigma_{oo}^{-1}(\beta_2, \dots, \beta_K)^T$ , which is most likely to be different from the true effect size  $\beta_1$ , and the variance of the imputed data is  $\Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}$ . Hence, the expected value of the regression coefficient of the combined observed and imputed data,  $(1-r)\beta_1 + r\Sigma_{mo}\Sigma_{oo}^{-1}(\beta_2, \dots, \beta_K)^T$ , is biased and its estimate can not be used as an inference of the true SNV effect  $\beta_1$ . In addition, GWAS results often suffer from “winner’s curse”, a phenomenon where the estimate of the genetic effect tends to have an upward bias, which requires a replication study to estimate the true effect size. We recommend to use the P-value from the association analysis of the combined observed and imputed data as the final P-value of the SNV and use the effect size estimate from the analysis of the observed data alone as the inference of the SNV effect. Therefore, we can achieve an improved P-value and an estimate of the effect size which is not biased due to

the imputation.

### 3.2.7 Factors Influencing Imputation Accuracy

We consider using the correlation and mean square error (MSE) between the imputed and the true unobserved phenotype values as measures of imputation accuracy. Specifically, for unrelated study subjects, we have

$$\begin{aligned}
cor(y_{1,i}, \hat{y}_{1,i}) &= \frac{cov(y_{1,i}, \hat{y}_{1,i})}{\sqrt{var(y_{1,i}) \cdot var(\hat{y}_{1,i})}} \\
&= \frac{cov(y_{1,i}, \bar{y}_1 + \Sigma_{mo} \Sigma_{oo}^{-1} (y_{2,i} - \bar{y}_2, \dots, y_{K,i} - \bar{y}_K)^T)}{\sqrt{1 \cdot \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om}}} \\
&= \frac{cov(y_{1,i}, \bar{y}_1) + \Sigma_{mo} \Sigma_{oo}^{-1} (cov(y_{1,i}, y_{2,i} - \bar{y}_2), \dots, cov(y_{1,i}, y_{K,i} - \bar{y}_K))^T}{\sqrt{1 \cdot \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om}}} \\
&= \frac{\Sigma_{mo} \Sigma_{oo}^{-1} (cov(y_1, y_2), \dots, cov(y_1, y_K))^T}{\sqrt{1 \cdot \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om}}} \\
&= \sqrt{\Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om}}
\end{aligned} \tag{3.13}$$

and

$$\begin{aligned}
MSE &= \sum_{i=n(1-r)+1}^n (\bar{y}_1 + \Sigma_{mo}\Sigma_{oo}^{-1}(y_{2,i} - \bar{y}_2, \dots, y_{K,i} - \bar{y}_K)^T - y_{1,i})^2 / nr \\
&= \sum_{i=n(1-r)+1}^n [(y_{1,i} - \bar{y}_1)^2 + (\Sigma_{mo}\Sigma_{oo}^{-1}(y_{2,i} - \bar{y}_2, \dots, y_{K,i} - \bar{y}_K)^T)^2 \\
&\quad - 2(y_{1,i} - \bar{y}_1)\Sigma_{mo}\Sigma_{oo}^{-1}(y_{2,i} - \bar{y}_2, \dots, y_{K,i} - \bar{y}_K)^T] / nr \\
&= var(y_1) + \Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{oo}\Sigma_{oo}^{-1}\Sigma_{om} - 2\Sigma_{mo}\Sigma_{oo}^{-1}(cov(y_1, y_2), \dots, cov(y_1, y_K))^T \\
&= var(y_1) - \Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}
\end{aligned} \tag{3.14}$$

The above derivation shows that the MSE is determined by the variance and covariance of the  $K$  collected phenotypes. The same conclusion holds for the correlation as well. In addition, the percentage of missing data can also affect the imputation accuracy in a family study. Unlike a population-based study in which only the missing sample's collected phenotypes are leveraged, the relatives of the missing sample also contribute to the imputed values in a family study. A large percentage of missing data can decrease the number of relatives contributing, therefore percentage of missing data is a factor of imputation accuracy in family studies.

### 3.2.8 Simulation Evaluation for Single Variant Test

The goals of the simulation study are 1) to examine and compare our approximated NCP derivation and the NCP derived under NMM for the single variant test using unrelated

samples; 2) to evaluate imputation power and type-I error rate under different conditions using family data and 3) to examine imputation accuracy with and without using family structure in the imputation.

To achieve the first goal, a locus with  $MAF = 0.2$  is generated using a binomial distribution. We simulate three phenotypes for each of 2000 unrelated individuals. Because the test statistic and NCP (under NMM assumption) in the original PhenIMP paper were derived with standardized data, we also standardize our simulated phenotypes and genotypes for comparison. We assume that the polygenic and environmental covariance matrices are

$$\Sigma_A = \Sigma_E = \begin{pmatrix} 0.5 & \pm 0.25 & \pm 0.25 \\ \pm 0.25 & 0.5 & \pm 0.25 \\ \pm 0.25 & \pm 0.25 & 0.5 \end{pmatrix} \quad (3.15)$$

which imply that each phenotype has a heritability of 50% and the pair-wise correlations between phenotypes are set to  $\pm 0.5$ . We assume that the first phenotype has a missing percentage of 20%, 50% or 80%, and the second and third phenotypes are complete. The missing values in the first phenotype are imputed using the second phenotype alone, the third phenotype alone, and the second and the third phenotypes together. The phenotypic variances of the three phenotypes explained by the single variant ( $R^2$ ) are set to 2%, 1%

and 0.8%, respectively, and the effect sizes are determined by  $\beta = \sqrt{\frac{R^2}{2p(1-p)}}$ .

From the above NCP derivation, it is easy to see that the product of the expected effect size of the phenotypes and the pair-wise correlations is the factor affecting the statistical power of the single variant test. Therefore, we fix the direction of the effect sizes (all  $> 0$ ) and vary the sign of the pair-wise correlations in order to examine every possible combination in the simulation. The analysis to verify the NCP derivation is implemented in the R package seqMeta.

In order to evaluate our family-informed method, we generate 500 nuclear families consisting of two parents and two children for a total sample size of 2000. For type-I error rate evaluation, we generate 50000 SNVs with MAF sampled from a uniform distribution between 0.05 to 0.5. The genotypes for the parents are simulated first, then we randomly assign one allele from each parent to generate the genotypes for their offspring. The phenotype values are simulated from the MVN distribution previously described. Note that the kinship matrix  $\Phi$  is computed from the pedigree and is not equal to the identity matrix. We assign different constant mean levels for different phenotypes, so none of the 50000 SNVs are associated with either the phenotype of interest (Phe1) or the phenotypes we use to impute (Phe2 and Phe3). Missing percentage is set to 20%, 50% and 80%. The simulations are repeated 3000 times for a total of 150 million SNVs and we examine

type-I error rate at significance levels of 0.05,  $1 \times 10^{-3}$  and  $1 \times 10^{-5}$ .

To evaluate power, we use the same family settings (500 nuclear families for a total of 2000 individuals). A single variant is tested and the phenotype values are generated from the MVN distribution using the same kinship matrix  $\Phi$  as above. The mean level of the MVN distribution is computed using the same settings (MAF,  $\beta$ ,  $R^2$  and correlation between phenotypes) as the simulation for NCP examination. The association test for type-I error rate and power evaluation is performed using a linear mixed effects model in the software EMMA.

The third goal of the simulation is to examine imputation accuracy under different conditions. We use family data that are generated for power evaluation. The phenotype pair-wise correlations are restricted to be positive. The missing phenotype values are imputed using our family-informed method and PhenIMP which ignores the family structure. We also examine the effect of different missing percentage (20%, 50% and 80%) on imputation accuracy. Both correlation and MSE between the true and imputed values are computed as measurements of imputation accuracy.

### **3.2.9 Application of 2 Hour Glucose in FHS**

FHS is a longitudinal study initiated in the year of 1948. It has made significant contribution in identifying the risk factors for CVD, as well as other diseases including

Type-2 Diabetes (T2D). FHS consists of three generations and 14428 individuals from Framingham, MA: the original cohort, the offspring cohort and the third generation cohort (Gen3). Fasting glucose (FG) and 2 hour glucose (2hrglu) were obtained in FHS Offspring and Gen3 Cohorts with genotypes available in SNP Health Association Resource (SHARe). These two phenotypes were collected in exam 5 for the Offspring Cohort and exam 2 in the Gen3 Cohort. There are 173 individuals with missing 2 hour glucose values and the FG values are complete. We exclude individuals with missing T2D status or T2D cases. A linear mixed effects model with random intercept to account for familial correlation is used with adjustment for age, age<sup>2</sup>, sex, BMI and cohort indicator to test the association between five known loci (rs1260326, rs2877716, rs12243326 and rs17271305 and rs10423928) and 2 hour glucose [40]. We impute the 173 missing 2 hour glucose values using FG and test the association in the dataset of all observed 2 hour glucose values, and the combined observed and imputed dataset. We also set different percentages of individuals to missing and impute them using FG to examine the performance of our method.

### **3.3 Results**

We first examine the NCP approximation for single variant test by comparing the empirical power with the approximated theoretical power (Table 3.1, 3.2, 3.3). Two datasets are used in the evaluation: the imputed values alone (Imp), and the combined imputed and observed data (Imp+Obs). When using one phenotype (either phenotype 2 or

3) to impute the missing phenotype 1 values, we vary the direction of the correlation (positive and negative) between the two phenotypes. And when using both phenotypes in the imputation, we examine all 8 combinations of different pair-wise correlation directions. Note that because the data are standardized, the NCP of the pooled data is equivalent to the NCP of the Stouffer's signed Z-score method with optimal weights. Under the assumption of NMM, the NCP is determined by the effect size of phenotype 1 and the correlation between the phenotypes, so we expect to see the same NCP for using phenotypes 2 and 3 because the correlation between phenotype 2 and phenotype 1 is equal to the correlation between phenotype 3 and phenotype 1.

The results show that the empirical power is very similar to the theoretical power computed using our theoretical approximation, with random fluctuation in an acceptable range in all situations. The theoretical power computed using NMM does not take the direction of pair-wise phenotype correlations into consideration and hence can not predict the statistical power correctly.

Another interesting finding is that the power of Imp+Obs is even lower than the power of Imp in some cases. For example, in the last row of Table 3.1 (pair-wise correlation between phenotypes is labeled as “- - -”), the power of the imputed data is 22.2%, but the power for the combined imputed and observed data is only 18%. This is because the effect size of the observed data is in an opposite direction of the effect size of the imputed data

which is determined by the pair-wise correlation between phenotypes and the direction of the effect sizes of the phenotypes included in the imputation (phenotypes 2 and 3 in this example). The power of Imp+Obs is lower than the power of Imp because the effect sizes of Imp and Obs compensate for each other in Imp+Obs.

We examine the type-I error rate of family data in the imputed only and the combined imputed and observed datasets under different missing percentages. All of the type-I error rates are correctly controlled (Table 3.4).

Next, we compare the empirical power of the combined and imputed family data vs. the power of the incomplete data (Table 3.5). When using one additional phenotype, either Phe2 or Phe3, imputation can boost power when the sign of  $\varphi_{12}$  ( $\varphi_{13}$ ) is the same as the sign of  $\beta_1 \times \beta_2$  ( $\beta_1 \times \beta_3$ ), and it reduces power when they are not the same. When imputing the missing values from two additional phenotypes, the power is determined by the signs of  $\beta_1 \times \beta_2$  and  $\beta_1 \times \beta_3$ . If  $\varphi_{12}$  has the same sign as  $\beta_1 \times \beta_2$  and  $\varphi_{13}$  has the same sign as  $\beta_1 \times \beta_3$ , we will see a power boost, otherwise there is a loss in power. These conclusions also hold in simulations with unrelated samples and can be verified using the NCP derived in the theoretical power approximation.

We then evaluate the effect of incorporating family structure information in the imputation on imputation accuracy when samples are related (Tables 3.6, 3.7). As expected, the correlation is higher and the MSE is lower when taking the family structure into account. It indicates that the information contained in families should be leveraged in the

Table 3.1: Power evaluation results of simulated data with 20% missing percentage

$P=1 \times 10^{-5}$			Empirical Power		Theoretical Power (NCP)								
	$\phi_{12}$	$\phi_{13}$	$\phi_{23}$	Imp	Imp+Obs	Imp	NMM	Imp	Imp+Obs	Imp	Imp+Obs		
Phe2	+			0.8%	93.9%	0.1%	(1.41)	92.1%	(5.83)	0.8%	(2.00)	94.0%	(5.97)
	-			0.8%	69.9%					0.8%	(-2.00)	72.1%	(5.00)
Phe3		+		0.4%	92.3%	0.1%	(1.41)	92.1%	(5.83)	0.4%	(1.79)	93.4%	(5.92)
		-		0.2%	73.2%					0.4%	(-1.79)	73.8%	(5.05)
Phe2 + Phe3	+	+	+	1.4%	94.7%	0.3%	(1.63)	92.9%	(5.89)	1.3%	(2.19)	94.8%	(6.04)
	+	+	-	21.8%	98.8%	5.6%	(2.83)	97.2%	(6.32)	26.5%	(3.79)	99.0%	(6.75)
	+	-	+	0%	76.4%	5.6%	(2.83)	97.2%	(6.32)	0%	(0.21)	76.9%	(5.15)
	+	-	-	0%	84.0%	0.3%	(1.63)	92.9%	(5.89)	0%	(0.12)	85.4%	(5.47)
	-	+	+	0%	69.5%	5.6%	(2.83)	97.2%	(6.32)	0%	(-0.21)	70.8%	(4.97)
	-	+	-	0%	83.3%	0.3%	(1.63)	92.9%	(5.89)	0%	(-0.12)	83.7%	(5.40)
	-	-	+	1.6%	64.8%	0.3%	(1.63)	92.9%	(5.89)	1.3%	(-2.19)	65.9%	(4.82)
	-	-	-	22.2%	18.0%	5.6%	(2.83)	97.2%	(6.32)	26.5%	(-3.79)	14.6%	(3.37)

'Phe2', 'Phe3', 'Phe2 + Phe3' represent the results of imputation using phenotype 2 alone, phenotype 3 alone, and phenotypes 2 and 3 together.  $\phi_{ij}$  represents the correlation between phenotypes  $i$  and  $j$ . Imp: imputed data only; Imp+Obs: combined imputed and observed data.

Table 3.2: Power evaluation results of simulated data with 50% missing percentage

$P = 1 \times 10^{-5}$				Theoretical Power (NCP)					
Empirical Power				NMM					
	$\varphi_{12}$	$\varphi_{13}$	$\varphi_{23}$	Imp	Imp+Obs	Imp	Imp+Obs	Imp	Imp+Obs
Phe2	+			9.6%	83.0%	1.5% (2.24)	72.0% (5.00)	10.5% (3.16)	84.1% (5.41)
	-			9.6%	3.6%			10.5% (-3.16)	3.4% (2.59)
Phe3		+		4.8%	79.8%	1.5% (2.24)	72.0% (5.00)	5.6% (2.83)	80.2% (5.26)
		-		6.2%	4.4%			5.6% (-2.83)	4.6% (2.74)
Phe2 + Phe3	+	+	+	12.8%	86.6%	3.3% (2.58)	77.2% (5.16)	16.9% (3.46)	88.2% (5.60)
	+	+	-	93.6%	99.8%	52.2% (4.47)	97.2% (6.32)	94.2% (5.99)	99.9% (7.40)
	+	-	+	0%	15.4%	52.2% (4.47)	97.2% (6.32)	0% (0.33)	15.4% (3.40)
	+	-	-	0%	37.3%	3.3% (2.58)	77.2% (5.16)	0% (0.19)	32.7% (3.97)
	-	+	+	0%	7.0%	52.2% (4.47)	97.2% (6.32)	0% (-0.33)	6.8% (2.93)
	-	+	-	0%	28.3%	3.3% (2.58)	77.2% (5.16)	0% (-0.19)	26.1% (3.78)
	-	-	+	13.6%	1.6%	3.3% (2.58)	77.2% (5.16)	16.9% (-3.46)	1.1% (2.14)
	-	-	-	92.6%	0%	52.2% (4.47)	97.2% (6.32)	94.2% (-5.99)	0% (-1.07)

Table 3.3: Power evaluation results of simulated data with 80% missing percentage

$P=1 \times 10^{-5}$				Theoretical Power (NCP)									
				Empirical Power			NMM			Approximated			
	$\varphi_{12}$	$\varphi_{13}$	$\varphi_{23}$	Imp	Imp+Obs	Imp	Imp+Obs	Imp	Imp+Obs	Imp	Imp+Obs	Imp	Imp+Obs
Phe2	+			32.7%	63.9%	3.8% (2.65)	47.7% (4.36)	33.8% (4.00)	66.0% (4.83)	33.8% (4.00)	66.0% (4.83)	33.8% (4.00)	66.0% (4.83)
	-			32.9%	0%			33.8% (-4.00)	0% (-0.83)			33.8% (-4.00)	0% (-0.83)
Phe3		+		22.2%	55.9%	3.8% (2.65)	47.7% (4.36)	20.1% (3.58)	54.5% (4.53)	20.1% (3.58)	54.5% (4.53)	20.1% (-3.58)	0% (-0.53)
		-		20.5%	0%			20.1% (-3.58)	0% (-0.53)			20.1% (-3.58)	0% (-0.53)
Phe2 + Phe3	+	+	+	48.1%	77.0%	8.7% (3.06)	58.0% (4.62)	48.3% (4.37)	77.1% (5.16)	48.3% (4.37)	77.1% (5.16)	48.3% (4.37)	77.1% (5.16)
	+	+	-	99.6%	100%	80.9% (5.29)	97.2% (6.32)	100.0% (7.58)	100.0% (8.04)	100.0% (7.58)	100.0% (8.04)	100.0% (7.58)	100.0% (8.04)
	+	-	+	0%	0.2%	80.9% (5.29)	97.2% (6.32)	0% (0.42)	0.3% (1.64)	0% (0.42)	0.3% (1.64)	0% (0.42)	0.3% (1.64)
	+	-	-	0%	1.6%	8.7% (3.06)	58.0% (4.62)	0% (0.24)	0.9% (2.04)	0% (0.24)	0.9% (2.04)	0% (0.24)	0.9% (2.04)
	-	+	+	0%	0%	80.9% (5.29)	97.2% (6.32)	0% (-0.42)	0% (0.89)	0% (-0.42)	0% (0.89)	0% (-0.42)	0% (0.89)
	-	+	-	0%	2.2%	8.7% (3.06)	58.0% (4.62)	0% (-0.24)	0.3% (1.67)	0% (-0.24)	0.3% (1.67)	0% (-0.24)	0.3% (1.67)
	-	-	+	46.3%	0.2%	8.7% (3.06)	58.0% (4.62)	48.3% (-4.37)	0.2% (-1.46)	48.3% (-4.37)	0.2% (-1.46)	48.3% (-4.37)	0.2% (-1.46)
	-	-	-	100.0%	80.2%	80.9% (5.29)	97.2% (6.32)	100.0% (-7.58)	86.3% (-5.5)	100.0% (-7.58)	86.3% (-5.5)	100.0% (-7.58)	86.3% (-5.5)

Table 3.4: Relative Type-I error rate of family data

$\alpha$ Level	Missing Percentage			Type-I error rate					
				0.05		$10^{-3}$		$10^{-5}$	
	Imp	Imp+Obs		Imp	Imp+Obs	Imp	Imp+Obs	Imp	Imp+Obs
Phe2	20%	1.00	1.00	1.00	1.00	1.00	1.01	1.06	1.01
	50%	1.00	1.00	1.00	1.00	1.00	1.00	1.01	0.98
	80%	1.00	1.00	1.00	1.00	0.99	0.98	0.99	0.99
Phe2 + Phe3	20%	1.00	1.00	1.00	1.00	1.00	1.02	0.97	0.97
	50%	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.95
	80%	1.00	1.00	1.00	1.00	0.98	0.99	1.01	1.02

Relative type-I error rate is the ratio between the empirical type-I error rate obtained from simulation and the expected type-I error rate, which is the significance level.

Table 3.5: Power evaluation for the combined observed and imputed family data

	Phen. Corr.			Missing Percentage		
	$\varphi_{12}$	$\varphi_{13}$	$\varphi_{23}$	20%	50%	80%
$\mathbf{P= 1 \times 10^{-5}}$						
Complete				95.9%	96.6%	96.2%
Incomplete				87.6%	49.0%	5.4%
Phe2	+			91.3%	77.7%	51.7%
	-			73.6%	12.4%	0%
Phe3		+		90.7%	75.6%	43.3%
		-		70.4%	11.0%	1.4%
Phe2 + Phe3	+	+	+	92.2%	82.8%	61.4%
	+	+	-	97.2%	99.2%	100%
	+	-	+	84.8%	35.7%	1.0%
	+	-	-	83.0%	40.1%	3.6%
	-	+	+	83.2%	25.5%	0%
	-	+	-	82.4%	38.0%	4.6%
	-	-	+	67.6%	8.8%	2.0%
	-	-	-	55.4%	0%	2.3%

imputation. As expected, analyses with families of 2 parents and 4 offsprings have a better imputation accuracy than families with 2 parents and 2 offsprings. This is because families with more relatives provide more information that can be exploited in the imputation. Missing percentage affects the imputation accuracy of family data because fewer family members contribute to the imputation under a higher missing percentage. In addition, the effect size of the SNV on the phenotype does not affect the imputation accuracy. This is verified by comparing correlation and MSE between using Phe2 and Phe3, where the SNV has a bigger effect on Phe2.

The FHS glyceemic dataset has 5,830 individuals from the Offspring and Gen3 Cohorts. FG is observed on every individual, while 2 hour glucose has 173 missing values. The

Table 3.6: Imputation accuracy of family data (2 parents + 2 offsprings)

	Missing Percentage	Family structure			
		No	Yes	No	Yes
		Correlation		MSE	
Phe2	20%	0.50	0.58	0.75	0.67
	50%	0.50	0.55	0.75	0.70
	80%	0.50	0.52	0.75	0.73
Phe3	20%	0.51	0.58	0.75	0.66
	50%	0.50	0.55	0.75	0.70
	80%	0.51	0.52	0.74	0.73
Phe2 + Phe3	20%	0.58	0.64	0.66	0.59
	50%	0.58	0.62	0.66	0.62
	80%	0.58	0.60	0.67	0.65

‘No’: ignoring family structure in the imputation; ‘Yes’: including family structure in the imputation.

Table 3.7: Imputation accuracy of family data (2 parents + 4 offsprings)

	Missing Percentage	Family structure			
		No	Yes	No	Yes
		Correlation		MSE	
Phe2	20%	0.50	0.60	0.75	0.63
	50%	0.50	0.57	0.75	0.67
	80%	0.50	0.53	0.75	0.72
Phe3	20%	0.50	0.60	0.74	0.63
	50%	0.51	0.57	0.74	0.67
	80%	0.51	0.54	0.74	0.71
Phe2 + Phe3	20%	0.58	0.66	0.66	0.56
	50%	0.58	0.64	0.66	0.60
	80%	0.58	0.61	0.66	0.64

‘No’: ignoring family structure in the imputation; ‘Yes’: including family structure in the imputation.

Table 3.8: Association tests results of FG and 2 hour glucose in FHS

SNV	FG			2hrglu		
	$\beta$	SE	P	$\beta$	SE	P
rs1260326	-0.632	0.158	$6.3 \times 10^{-5}$	0.920	0.503	0.068
<b>rs2877716</b>	-0.521	0.186	$5.2 \times 10^{-3}$	-1.659	0.594	$5.2 \times 10^{-3}$
<b>rs12243326</b>	0.418	0.173	0.016	1.658	0.552	$2.7 \times 10^{-3}$
rs17271305	-0.133	0.162	0.41	1.020	0.516	0.048
rs10423928	-0.284	0.199	0.63	2.910	0.634	$4.5 \times 10^{-6}$

Bold: expected to see an increase in power.

sample correlation between FG and 2 hour glucose is 0.39. We select 5 loci previously identified to be associated with 2 hour glucose [40]. We first test the association of these 5 loci in our data. Four of the 5 loci (rs2877716, rs12243326, rs17271305 and rs10423928) are significantly associated with 2 hour glucose, with a P-value less than 0.05, and 1 locus (rs1260326) is just slightly above the 0.05 threshold (Table 3.8). Based on the approximated NCP derivation and the simulation results above, we expect a more significant association test result on rs2877716 and rs12243326 after imputation because these 2 loci are associated with FG and the effect sizes are in the same direction as their effect sizes on 2 hour glucose. For the other 3 loci, their effect sizes on FG and 2 hour glucose are in different directions, hence we do not expect to see a more significant result by imputing from FG.

We refer to the dataset of all 5,830 individuals as “All” and the dataset of the 5,657 individuals with no missing FG and 2 hour glucose as “Obs”. We first assess the association for the 5 loci in “Obs”. Then the 173 missing values of 2 hour glucose in “All” are imputed using FG and the association analysis is repeated. To fully examine our

method, we also randomly set 3%, 10%, 25%, 50% and 80% 2 hour glucose values in “Obs” to missing, then impute the missing values using FG. We then test the 5 loci in both the incomplete dataset in which missing values are removed, referred to as “(1-K%) Obs”, and the combined incomplete and imputed dataset, referred to as “(1-K%) Obs + K% Imp”, where K% is the missing percentage. We repeat this process 200 times and report the median of the P-values from the 200 iterations.

For rs1260326, rs17271305 and rs10423928, imputation from FG does not improve the association test results: P-values from “Obs” are more significant than “All”, and the incomplete dataset “(1-K%) Obs” outperforms the combined dataset “(1-K%) Obs + K% Imp” for all 5 missing percentage scenarios (Table 3.9). This is not surprising because their effects on FG and 2 hour glucose are in different directions. The imputation improves significance for rs2877716 as expected. For rs12243326, we find a more significant P-value when the missing percentage is over 25%. Although the imputation does not improve the significance of P-values when the missing percentages are below 10%, P-values before and after imputation are the same or very close. As expected, a larger missing percentage can cause the loss in power (a less significant P-value) in both incomplete and combined datasets.

Table 3.9: Median P-values from FHS 2 hour glucose data

<b>Dataset</b>	<b>Sample Size</b>	<b>rs1260326</b>	<b>rs2877716</b>	<b>rs12243326</b>	<b>rs17271305</b>	<b>rs10423928</b>
Obs	5657	0.068	$5.2 \times 10^{-3}$	$2.7 \times 10^{-3}$	0.048	$4.5 \times 10^{-6}$
All	5830	0.094	$4.9 \times 10^{-3}$	$2.8 \times 10^{-3}$	0.059	$4.8 \times 10^{-6}$
97% Obs + 3% Imp	5657	0.081	$5.4 \times 10^{-3}$	$2.8 \times 10^{-3}$	0.049	$6.7 \times 10^{-6}$
97% Obs	5488	0.072	$6.1 \times 10^{-3}$	$2.8 \times 10^{-3}$	0.046	$5.6 \times 10^{-6}$
90% Obs + 10% Imp	5657	0.11	$5.5 \times 10^{-3}$	$4.4 \times 10^{-3}$	0.061	$2.5 \times 10^{-5}$
90% Obs	5092	0.087	$8.1 \times 10^{-3}$	$4.4 \times 10^{-3}$	0.054	$1.0 \times 10^{-5}$
75% Obs + 25% Imp	5657	0.25	$4.9 \times 10^{-3}$	$4.9 \times 10^{-3}$	0.080	$2.3 \times 10^{-4}$
75% Obs	4243	0.12	0.011	$6.9 \times 10^{-3}$	0.059	$3.1 \times 10^{-5}$
50% Obs + 50% Imp	5657	0.56	$9.3 \times 10^{-3}$	0.010	0.22	$8.1 \times 10^{-3}$
50% Obs	2829	0.28	0.037	0.032	0.10	$6.3 \times 10^{-4}$
20% Obs + 80% Imp	5657	0.18	0.016	0.020	0.54	0.49
20% Obs	1132	0.49	0.25	0.12	0.36	0.046

### 3.4 Discussion

We extended PhenIMP to leverage information contained not only in correlated phenotypes but also family structure. We showed that taking the family structure into consideration could improve imputation accuracy. In addition, we investigated the situations where our method can increase power and the situations we do not expect power gains from imputation. We also derived the theoretical NCP for PhenIMP because the NCP based on NMM can not accurately predict power in the single variant association test. In the analysis of the combined observed and imputed data, we found that the pooled analysis is equivalent to the meta-analysis of Z score when the data are standardized.

Our method uses the same MVN distribution model as PhenIMP and PHENIX. These three methods also share the same underlying assumption of pleiotropy and hence may lose power when this assumption does not hold [38]. We performed the first investigation of imputation accuracy and power, to our knowledge, by extensive simulations and we determined factors that could affect imputation accuracy and power. Our conclusions can also be generalized to PhenIMP and PHENIX because the three methods have the same assumption of MVN distribution.

The model we propose can impute missing values on multiple phenotypes at once, hence we do not need to repeat the same imputation process for each missing phenotype. We estimate the genetic and environmental variance components through MLE, without the

need to collect pilot data prior to the analysis. However, we acknowledge that when the number of observed individuals is too small (high missing percentage), it might be better to use a pilot dataset to estimate the pair-wise phenotype correlations because the MLE may provide a poor estimate of the true value. Through simulation, we showed that MLE worked well even under an 80% missing percentage.

Intuitively, with more phenotypes being used in the imputation, we would expect a better imputation accuracy because more phenotypes contain more information. However, based on our simulation results, the number of phenotypes and the pair-wise correlation between phenotypes are the most important factors that determine the accuracy of the imputation. The sign of effect sizes and the correlation between phenotypes determine whether or not the power can be improved after imputation. Including phenotypes with different directions of the SNV effect might decrease power. We suggest to start with one or two additional phenotypes because it's easy to keep track of the change in NCP when the number of phenotypes is small.

The analysis of the combined observed and imputed data needs to be performed cautiously. Because the imputed data have a biased effect size estimate and a different variance than the observed data, the effect size estimates from the combined dataset are not accurate. Here we suggest to use the P-value from the combined dataset since the imputation can improve power to detect the association and use the effect size estimate

from the observed dataset because it is not subject to the bias due to imputation.

## Chapter 4 Extension of a Phenotype Imputation Approach to Variant-Set Tests

### 4.1 Introduction

Common variants found to be associated with a trait and identified by GWAS can only explain a small proportion of heritability. For example, although Alzheimer's Disease has a heritability of  $h^2 = 0.58 - 0.79$  [41], not much heritability can be explained by the major genetic risk factor *APOE* [42, 43] gene and other loci identified in European ancestry individuals [44, 45, 46]. It is still not clear how to explain the missing heritability that can not be accounted for by common variants detected so far. With the huge fall in genome sequencing costs, more focus has been placed on rare variants to try to explain the missing heritability. Because single-variant tests for rare variants remain challenging due to the lack of power to detect associations, several variant-set tests have been developed to address the issue of low power of rare variant tests.

One class of the variant-set test is burden tests. In burden tests, the genotypes of single variants in the region are collapsed into an aggregated score. A regression model is then fitted with the aggregated score as the independent variable and the phenotype as the dependent variable. Burden tests are most powerful when all variants within the region have the same direction of effect and the proportion of causal SNVs is high. Alternatively, SKAT uses a variance component model and retains power when the variants have different directions of effect. It has been proven to be more powerful than burden tests when variants have different directions of effect. The SKAT statistic is equivalent to the weighted sum of the score statistic for single variants, which is very useful for

meta-analysis.

In the previous chapter, we developed a family-informed phenotype imputation approach and we showed that our approach can improve statistical power in single variant association tests under certain situations. But the performance of our method in variant-set tests has not been evaluated. Intuitively, the situation is more complicated because different SNVs in the same region could affect the phenotype in different directions and not all SNVs are causal. Hence, there is a need to further investigate the performance of our phenotype imputation approach in variant-set tests.

In this chapter, we first derive an approximation to the theoretical power for two variant-set tests, burden test and SKAT, in unrelated samples. Then we use simulation to evaluate our theoretical power approximation, examine type-I error rate and investigate situations where our method can boost power and situations where imputation can not improve statistical significance in unrelated and family samples, respectively. Lastly, we use a real data example of FHS to validate our findings in simulation study.

## **4.2 Methods**

### **4.2.1 Power of Burden Test**

We derive an analytical approximation to compute power for two popular variant-set tests, burden test and SKAT, when the study subjects are unrelated. We focus on the situation in

which the combined observed and imputed data are analyzed jointly. The derivation using the imputed data alone is similar and requires fewer steps.

In a burden test, one way to compute power theoretically is consider the burden score as a single variant. When the aggregated variants have a cumulative minor allele frequency (cMAF), say 5%, then the power is equivalent to a single variant test with the same MAF. So we can directly use the non-centrality parameter (NCP) for single variant test we derived in chapter 3 to compute power for burden test. However, using this method, we assume that all single variants within the gene have the same effect size, which is not true in most cases. Here we show an analytical approach to compute power which allows different effect size for each variant.

We assume  $y$  is a length  $n$  vector of the observed and imputed phenotype 1 values, and we divide  $y$  into two parts:  $y_o$  the length  $n(1-r)$  vector of the  $n(1-r)$  samples with the observed phenotype 1 values and  $y_m$  the length  $nr$  vector of the  $nr$  imputed samples. Similarly, we assume  $s$  is a vector of the aggregated burden for  $n$  individuals and we divide  $s$  into  $s_o$  the burden scores for observed individuals and  $s_m$  the burden scores for imputed individuals. In addition, we assume  $\phi$  is a  $l \times l$  correlation matrix of the  $l$  SNVs within the gene, where  $\phi_{ij}$  describes the correlation between SNV  $i$  and SNV  $j$ , and  $w_j$  is the weight for SNV  $j$ . The burden score vector  $s_o$  and  $s_m$  can be computed as  $\sum_{i=1}^l w_i x_{o,i}$  and  $\sum_{i=1}^l w_i x_{m,i}$ , where  $x_{o,i}$  and  $x_{m,i}$  are the genotype vector for SNV  $i$  on the  $n(1-r)$

observed and  $nr$  imputed samples, respectively.

$$\begin{aligned}
E(\hat{\beta}) &= E((s^T s)^{-1} s^T y) \\
&= E\left(\frac{s^T y}{s^T s}\right) \\
&= E\left(\frac{s_o^T y_o + s_m^T y_m}{s_o^T s_o + s_m^T s_m}\right) \\
&= \frac{\sum_{i=1}^l w_i x_{o,i}^T [x_{o,1}, \dots, x_{o,l}] \beta_1 + \sum_{i=1}^l w_i x_{m,i}^T [x_{m,1}, \dots, x_{m,l}] [\beta_2, \dots, \beta_K] \Sigma_{oo}^{-1} \Sigma_{om}}{\sum_{i=1}^l w_i x_{o,i}^T \sum_{i=1}^l w_i x_{o,i} + \sum_{i=1}^l w_i x_{m,i}^T \sum_{i=1}^l w_i x_{m,i}} \\
&= \frac{[\sum_{i=1}^l w_i x_{o,i}^T x_{o,1}, \dots, \sum_{i=1}^l w_i x_{o,i}^T x_{o,l}] \beta_1}{\sum_{i=1}^l w_i x_{o,i}^T \sum_{i=1}^l w_i x_{o,i} + \sum_{i=1}^l w_i x_{m,i}^T \sum_{i=1}^l w_i x_{m,i}} \\
&\quad + \frac{[\sum_{i=1}^l w_i x_{m,i}^T x_{m,1}, \dots, \sum_{i=1}^l w_i x_{m,i}^T x_{m,l}] [\beta_2, \dots, \beta_K] \Sigma_{oo}^{-1} \Sigma_{om}}{\sum_{i=1}^l w_i x_{o,i}^T \sum_{i=1}^l w_i x_{o,i} + \sum_{i=1}^l w_i x_{m,i}^T \sum_{i=1}^l w_i x_{m,i}} \\
&= \frac{n(1-r) [\sum_{i=1}^l w_i \phi_{i,1}, \dots, \sum_{i=1}^l w_i \phi_{i,l}] \beta_1}{n(1-r) \sum_{i=1}^l \sum_{j=1}^l w_i w_j \phi_{i,j} + nr \sum_{i=1}^l \sum_{j=1}^l w_i w_j \phi_{i,j}} \\
&\quad + \frac{nr [\sum_{i=1}^l w_i \phi_{i,1}, \dots, \sum_{i=1}^l w_i \phi_{i,l}] [\beta_2, \dots, \beta_K] \Sigma_{oo}^{-1} \Sigma_{om}}{n(1-r) \sum_{i=1}^l \sum_{j=1}^l w_i w_j \phi_{i,j} + nr \sum_{i=1}^l \sum_{j=1}^l w_i w_j \phi_{i,j}} \\
&= \frac{[\sum_{i=1}^l w_i \phi_{i,1}, \dots, \sum_{i=1}^l w_i \phi_{i,l}] ((1-r) \beta_1 + r [\beta_2, \dots, \beta_K] \Sigma_{oo}^{-1} \Sigma_{om})}{\sum_{i=1}^l \sum_{j=1}^l w_i w_j \phi_{i,j}}
\end{aligned} \tag{4.1}$$

The variance of the beta estimate can be computed as

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \sigma^2 (s^T s)^{-1} \\
&= ((1-r) + r \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om}) \left( n \sum_{i=1}^l \sum_{j=1}^l w_i w_j \phi_{i,j} \right)^{-1}
\end{aligned} \tag{4.2}$$

Hence the NCP of the test statistic of Burden test can be written as

$$\frac{\sqrt{n}[\sum_{i=1}^l w_i \phi_{i,1}, \dots, \sum_{i=1}^l w_i \phi_{i,l}][((1-r)\beta_1 + r[\beta_2, \dots, \beta_K] \Sigma_{oo}^{-1} \Sigma_{om})]}{\sqrt{((1-r) + r \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om}) \sum_{i=1}^l \sum_{j=1}^l w_i w_j \phi_{i,j}}} \quad (4.3)$$

#### 4.2.2 Power of SKAT

We assume that the linear model relating the variants in a region to the phenotype is  $y_i = \gamma_0 + C_i \gamma + G_i \beta$ , where  $y_i$  is the  $i$ th element of the phenotype vector  $y$ ,  $\gamma_0$  is the intercept,  $C_i$  is the  $i$ th row of the covariates matrix  $C$ ,  $\gamma$  is the vector of regression coefficients for the covariates,  $G_i$  is the  $i$ th row of the  $n \times l$  genotype matrix  $G$  of  $l$  SNVs and  $\beta$  is the length  $l$  vector of regression coefficients for the  $l$  SNVs.

The SKAT statistic has the form of  $Q = (y - \hat{\mu})' \mathbf{K} (y - \hat{\mu})$ , where  $\hat{\mu} = \hat{\gamma}_0 + C \hat{\gamma}$ ,  $\hat{\gamma}_0$  and  $\hat{\gamma}$  are the estimated regression coefficients under the null,  $\mathbf{K} = G W G'$  and  $W = \text{diag}(w(\hat{m}_1), \dots, w(\hat{m}_l))$  with  $\hat{m}_j$  being the MAF for SNV  $j$ .

Unlike SKAT with complete data, in which  $y - \hat{\mu}$  has  $I_{n \times n} \sigma_1^2$  as asymptotic covariance matrix, the combined observed and imputed phenotype vector  $y - \hat{\mu}$  follows MVN distribution with mean  $\mu_\beta = G \beta$  and an asymptotic covariance matrix

$$\Sigma = \begin{pmatrix} I_{n(1-r) \times n(1-r)} \sigma_1^2 & \mathbf{0}_{n(1-r) \times nr} \\ \mathbf{0}_{nr \times n(1-r)} & I_{nr \times nr} \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om} \end{pmatrix} \quad (4.4)$$

We first find the orthonormal matrix  $U$  that converts  $\Sigma^{1/2}\mathbf{K}\Sigma^{1/2}$  to its diagonal form  $\text{diag}(\lambda_1, \dots, \lambda_n) = U\Sigma^{1/2}\mathbf{K}\Sigma^{1/2}U^T$ . Then let  $Z = U\Sigma^{-1/2}(y - \hat{\mu})$  which follows a MVN distribution with mean  $\mu_z = U\Sigma^{-1/2}\mu_\beta$  and covariance matrix  $= I_{n \times n}$ . Now we can rewrite  $Q$  as a weighted sum of Chi-square variables

$$Q = (y - \hat{\mu})'\mathbf{K}(y - \hat{\mu}) = Z^T \text{diag}(\lambda_1, \dots, \lambda_n)Z = \sum_{j=1}^n \lambda_j \chi_1^2(\delta_j) \quad (4.5)$$

where  $\delta_j = \mu_{zj}^2$ ,  $\mu_{zj}$  is the  $j$ th element of  $\mu_z$ . Using the method proposed by Liu et al. [26], we approximate the distribution of  $Q$  using a non-central Chi-square variable. Following the procedures developed in SKAT, we need to compute

$$\begin{aligned} \sum_{j=1}^n \lambda_j^m &= \text{trace}((\mathbf{K}\Sigma)^m) \\ &= \text{trace}((GWG^T\Sigma)^m) \\ &= \text{trace}((WG^T\Sigma G)^m) \end{aligned} \quad (4.6)$$

and

$$\begin{aligned} \sum_{j=1}^n \lambda_j^m \delta_j &= \text{trace}(\mu_\beta^T (\mathbf{K}\Sigma)^{m-1} \mathbf{K} \mu_\beta) \\ &= \text{trace}(\mu_\beta^T (GWG^T\Sigma)^{m-1} GWG^T \mu_\beta) \\ &= \text{trace}(\mu_\beta^T G(WG^T\Sigma G)^{m-1} WG^T \mu_\beta) \end{aligned} \quad (4.7)$$

where  $\mu_\beta = \begin{pmatrix} G_1\beta_1 \\ G_2(\beta_2, \dots, \beta_K)\Sigma_{oo}^{-1}\Sigma_{om} \end{pmatrix}$ ,  $\beta_k$  is a length  $l$  vector which contains the effect sizes of  $l$  SNVs on phenotype  $k$ ,  $G_1$  is an  $n(1-r) \times l$  genotype matrix for the  $n(1-r)$  observed samples and  $G_2$  is an  $nr \times l$  genotype matrix for the  $nr$  samples with missing phenotype 1. To compute  $\sum_{j=1}^n \lambda_j^m$  and  $\sum_{j=1}^n \lambda_j^m \delta_j$ , we also need the following approximation

$$\begin{aligned}
G^T \Sigma G &= \begin{pmatrix} G_1^T & G_2^T \end{pmatrix} \begin{pmatrix} I_{n(1-r) \times n(1-r)} \sigma_1^2 & \mathbf{0}_{n(1-r) \times nr} \\ \mathbf{0}_{nr \times n(1-r)} & I_{nr \times nr} \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om} \end{pmatrix} \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \\
&= \begin{pmatrix} G_1^T \sigma_1^2 & \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om} G_2^T \end{pmatrix} \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \\
&= \sigma_1^2 G_1^T G_1 + \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om} G_2^T G_2 \\
&= \sigma_1^2 n(1-r) \phi + \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om} nr \phi
\end{aligned} \tag{4.8}$$

and

$$\begin{aligned}
G^T \mu_\beta &= \begin{pmatrix} G_1^T & G_2^T \end{pmatrix} \begin{pmatrix} G_1\beta_1 \\ G_2(\beta_2, \dots, \beta_K)\Sigma_{oo}^{-1}\Sigma_{om} \end{pmatrix} \\
&= G_1^T G_1 \beta_1 + G_2^T G_2 (\beta_2, \dots, \beta_K) \Sigma_{oo}^{-1} \Sigma_{om} \\
&= n(1-r) \phi \beta_1 + nr \phi (\beta_2, \dots, \beta_K) \Sigma_{oo}^{-1} \Sigma_{om}
\end{aligned} \tag{4.9}$$

where  $\phi$  is a  $l \times l$  matrix describing the covariance between the  $l$  SNVs. If the genotypes

are standardized, we can use the LD matrix as  $\phi$ . We then follow the steps described in [33] to approximate the theoretical power for SKAT.

### 4.2.3 Simulation

We use simulation to examine our theoretical power derivation for burden test and SKAT, and evaluate type-I error rate and power of these two variant-set tests for imputed phenotype values.

To generate single variants in the same region with LD, we first generate continuous latent variables from a MVN distribution, which has an order one autoregressive model with  $\rho = 0.8$  as the covariance matrix. Then for each variant, we assign 2 as the additive genotype to individuals with the latent variable below the  $p^2$  quantile, 0 when the latent variable is above the  $(1 - p)^2$  quantile and 1 for the rest, where  $p$  is the MAF of the variant.

In all simulations, we generate 2000 unrelated individuals and 2000 related individuals from 500 nuclear families with two parents and two children. When generating the related individuals, we first use the method described above to generate genotypes for all parents, then randomly assign one allele from each parent to generate the genotypes for the

children.

We first compare the empirical power from simulations to the approximated theoretical power. We simulate 5 single variants with MAF = 0.01, 0.015, 0.02, 0.025 and 0.03, respectively. The effect size of each single variant is determined by  $\sqrt{\frac{R^2}{2p(1-p)}}$ , where  $R^2$  is the proportion of phenotypic variance explained by the variant and  $p$  is the MAF. We set  $R^2$  to 0.1%, 0.1%, 0.5%, 0.2% and 0.2% for phenotype 1 (Phe1), 0.005%, 0, 0.3%, 0 and 0 for phenotype 2 (Phe2) and 0, 0, 0.1%, 0.2% and 0.1% for phenotype 3 (Phe3). All of the effect sizes are positive. We use the Wu weights ( $\text{dbeta}(1,25)$ ) in SKAT [27] and the Madsen and Browning weights ( $\sqrt{\frac{1}{maf(1-maf)}}$ ) in Burden test.

To evaluate type-I error rate and power under different conditions, we include 20 single variants with MAF randomly sampled from a uniform distribution between 0.002 and 0.05. Three phenotypes are simulated: one phenotype of interest (Phe1) and two additional phenotypes (Phe2 and Phe3). The pair-wise correlation between them is set to  $\pm 0.5$ . We vary the missing percentage from 20% to 50% and evaluate type-I error rate and power using unrelated and family data, respectively. In the type-I error rate evaluation, none of the 20 SNVs are associated with the phenotypes. When evaluating power, we set the absolute value of the effect size for each SNV to  $c|\log_{10}MAF|$ , where  $c = 0.2, 0.1$  and  $0.05$  for Phe1, Phe2 and Phe3, respectively. We also set the proportion of causal variants to 20%, 50% and 80% and vary the percentage of SNVs with the same directions of effect

from 100% to 50%.

#### **4.2.4 Application of 2 Hour Glucose in FHS**

The Framingham Heart study began in 1948 with participants from the town of Framingham, MA. It now has 14428 participants from 3 generations: Original, Offspring and the 3rd generation (Gen3). We apply our method on Offspring and Gen3 participants with glycemic traits available. We select 4 genes (*G6PC2*, *GCKR*, *GLP1R* and *VPS13C*) which contain at least one previously identified variant associated with 2 hour glucose or fasting glucose (FG) in [47, 48]. We perform SKAT (Wu weights) on individuals from the Offspring and Gen3 cohorts in FHS with adjustment of age, age<sup>2</sup>, BMI, cohort indicator and the top 10 PCs. Diabetic individuals are excluded from the analysis. The variant-set tests are restricted to nonsynonymous variants with MAF < 0.05 in exome-chip genotype set [18]. Among the 5627 individuals with available genotype data, 164 have missing 2 hour glucose values while FG is complete. Missing values are imputed using our family-informed imputation approach. We perform variant-set association tests on 5463 individuals with observed 2 hour glucose values, and the combined dataset of the 5463 observed and 164 imputed individuals. In addition, we randomly select 20% and 50% of the 5463 individuals and set their 2 hour glucose values to missing. We then impute them and test the association using the observed, and the combined observed and imputed datasets to evaluate the performance under different missing percentages.

Table 4.1: Power of combined observed and imputed data for unrelated samples with 20% missing percentage in variant-set tests.

<b>Power</b> $\alpha = 1 \times 10^{-5}$	<b>Phen Corr</b>			<b>Burden</b>		<b>SKAT</b>	
	$\varphi_{12}$	$\varphi_{13}$	$\varphi_{23}$	<b>Empirical</b>	<b>Theoretical</b>	<b>Empirical</b>	<b>Theoretical</b>
Phe2	+			95.2%	95.4%	96.5%	97.2%
	-			87.2%	86.8%	89.8%	91.6%
Phe3		+		96.0%	96.6%	97.1%	97.8%
		-		84.4%	85.0%	88.4%	89.7%
Phe2 + Phe3	+	+	+	96.2%	96.6%	97.1%	97.8%
	+	+	-	98.0%	98.6%	98.7%	99.1%
	+	-	+	78.8%	79.0%	82.7%	83.5%
	+	-	-	92.2%	92.4%	92.5%	93.4%
	-	+	+	87.2%	88.2%	88.8%	91.9%
	-	+	-	93.6%	93.6%	93.8%	95.2%
	-	-	+	86.4%	85.8%	86.0%	87.7%
	-	-	-	51.6%	53.0%	47.5%	49.7%

### 4.3 Results

We first examine our approach to compute theoretical power for burden test and SKAT under missing percentage of 20% (Table 4.1) and 50% (Table 4.2). The empirical power is similar to the approximated theoretical power in all cases indicating that our power approximation for SKAT and burden tests can be used to predict statistical power when designing a genetic association study.

Next, we examine the type-I error rate in unrelated (Table 4.3) and family data (Table 4.4). The missing phenotype values are imputed using either one (Phe2) or two (Phe2 + Phe3) additional phenotypes and we evaluate the type-I error rate in the imputed dataset, and the combined observed and imputed dataset, respectively, under  $\alpha$  level of 0.05,  $10^{-3}$  and

Table 4.2: Power of combined observed and imputed data for unrelated samples with 50% missing percentage in variant-set tests.

Power $\alpha = 1 \times 10^{-5}$	Phen Corr			Burden		SKAT	
	$\varphi_{12}$	$\varphi_{13}$	$\varphi_{23}$	Empirical	Theoretical	Empirical	Theoretical
Phe2	+			74.6%	76.5%	75.2%	78.7%
	-			24.2%	21.3%	22.8%	22.9%
Phe3		+		84.5%	82.6%	85.3%	87.1%
		-		15.4%	9.2%	15.6%	14.3%
Phe2 + Phe3	+	+	+	83.2%	83.0%	83.8%	86.3%
	+	+	-	97.7%	96.9%	97.9%	98.1%
	+	-	+	9.0%	5.6%	9.6%	9.0%
	+	-	-	36.8%	36.3%	37.6%	37.3%
	-	+	+	34.5%	26.2%	38.5%	38.4%
	-	+	-	50.6%	47.2%	50.6%	54.3%
	-	-	+	13.0%	8.5%	12.6%	9.3%
	-	-	-	0.2%	0%	0.2%	0%

$10^{-5}$ . All type-I error rates are correctly controlled for burden test and SKAT.

To investigate situations where our imputation approach can have a gain in power, we vary the sign of phenotype correlation, proportion of causal variants, percentage of variants with the same direction of effect and missing percentages in the simulation (Tables 4.5, 4.6, 4.7, 4.8). As expected, a high missing percentage can lead to a loss in power. Burden tests perform well when the effects of the variants are in the same direction and it almost has no power in the situation where 50% SNVs have a positive effect and 50% SNVs have a negative effect. Meanwhile, SKAT is powerful in both situations. The percentage of causal variants also affects power: the larger the percentage is, the higher the power we achieve. Our simulations assume that each variant affects the three phenotypes in the same direction, that is  $\beta_1, \beta_2, \beta_3$  all  $> 0$  or  $< 0$ , and the directions can be different for different

Table 4.3: Relative Type-I error rate of unrelated data in variant-set tests.

$\alpha$ Level	Test	r	Type-I error rate					
			0.05		$10^{-3}$		$10^{-5}$	
			Imp	Imp+Obs	Imp	Imp+Obs	Imp	Imp+Obs
Phe2	Burden	20%	1.00	1.00	1.00	1.01	1.02	1.01
		50%	0.97	0.99	0.98	1.00	0.96	0.98
	SKAT	20%	1.00	1.00	1.00	1.01	1.01	1.01
		50%	0.99	1.00	0.98	0.98	1.00	1.01
Phe2 + Phe3	Burden	20%	1.00	1.01	1.00	1.01	1.00	1.01
		50%	0.98	1.01	0.99	0.99	1.01	1.00
	SKAT	20%	1.00	1.00	1.00	1.01	1.01	1.01
		50%	1.00	0.98	1.02	1.01	1.01	1.00

Relative type-I error rate is the ratio between the empirical type-I error rate obtained from simulation and the expect type-I error rate, which is the significance level. r represents the missing percentage.

Table 4.4: Relative Type-I error rate of family data in variant-set tests.

$\alpha$ Level	Test	r	Type-I error rate					
			0.05		$10^{-3}$		$10^{-5}$	
			Imp	Imp+Obs	Imp	Imp+Obs	Imp	Imp+Obs
Phe2	Burden	20%	0.98	1.02	0.99	1.01	0.98	1.01
		50%	0.95	1.01	0.98	1.00	0.99	0.99
	SKAT	20%	0.98	1.00	0.98	0.99	1.00	1.00
		50%	0.99	1.00	1.01	1.01	0.97	0.98
Phe2 + Phe3	Burden	20%	0.93	0.99	0.92	0.97	0.89	0.97
		50%	0.96	1.01	1.02	0.98	0.97	0.97
	SKAT	20%	0.99	1.02	1.00	1.01	0.99	1.02
		50%	1.00	0.99	0.98	1.02	1.03	1.01

Relative type-I error rate is the ratio between the empirical type-I error rate obtained from simulation and the expect type-I error rate, which is the significance level. r represents the missing percentage.

variants (in the case of  $50\% > 0$  and  $50\% < 0$ ). Under this assumption, we obtain the same conclusions as for single variant tests in simulations with different signs of phenotype correlations. When using one additional phenotype, imputation can boost power if the sign of  $\varphi_{12}$  ( $\varphi_{13}$ ) is the same as the sign of  $\beta_1 \times \beta_2$  ( $\beta_1 \times \beta_3$ ). When using two additional phenotypes, power can be increased if  $\varphi_{12}$  has the same sign as  $\beta_1 \times \beta_2$  and  $\varphi_{13}$  has the same sign as  $\beta_1 \times \beta_3$ .

In our real data application of FHS glyceamic dataset, the estimated correlation between FG and 2 hour glucose is 0.39. First, we test the association between the 4 genes (*G6PC2*, *GCKR*, *GLP1R* and *VPS13C*) and FG or 2 hour glucose (Table 4.9). Two genes (*G6PC2* and *GLP1R*) are associated with FG, but none of them are significant in the test of 2 hour glucose. Hence, potentially we expect to see an increase of significance in genes *G6PC2* and *GLP1R*. Next, we impute the 164 missing 2 hour glucose values using their observed FG values. We refer to the 5463 individuals with observed 2 hour glucose values as “Obs” and 5627 individuals with the observed and imputed 2 hour glucose values as “All”. In the datasets where we randomly set samples from “Obs” missing, we use “(1-K%) Obs” to represent the incomplete data (missing values are removed), and “(1-K%) Obs + K% Imp” for the combined incomplete and imputed data, where K% is the missing percentage. Due to the very low missing percentage in “All” (3%), the significance of the association test results are similar between “All” and “Obs”. But when we use a higher missing percentage (20% and 50%), most of the P-values in “(1-K%) Obs + K% Imp” are lower than “(1-K%) Obs”. As expected, *GCKR* and *VPS13C* are still not significant after the



Table 4.6: Power of unrelated data (Imp+Obs) with 50% missing percentage in variant-set tests

Phen Direction Causal % Test	$\varphi_{12}$	$\varphi_{13}$	$\varphi_{23}$	Power (%)											
				20%				50%				80%			
				B	S	B	S	B	S	B	S	B	S	B	S
Complete				72.6	84.4	100	100	100	100	0.2	14.4	0.4	68.2	1.2	88.0
Incomplete				26.0	35.6	100	100	100	100	0	1.6	0.2	23.8	0.6	47.6
Phe2	+			39.0	54.0	100	100	100	100	0	3.6	0.2	35.8	0.6	63.8
	-			4.6	6.6	91.1	91.4	100	100	0	0.4	0	4.8	0.2	16.2
Phe3	+			25.8	37.0	99.8	100	100	100	0	1.8	0	23.0	0.2	50.6
	-			7.0	11.2	98.0	98.2	100	100	0	0.8	0.2	9.0	0.2	25.8
	+	+		37.6	48.4	100	100	100	100	0	3.2	0.2	30.4	0.4	59.6
	+	-		50.7	66.5	100	100	100	100	0	7.6	0	46.2	0.6	77.8
	+	-	+	14.2	20.8	99.8	99.8	100	100	0	0.4	0.2	14.8	0.4	37.4
	+	-	-	18.2	28.6	100	100	100	100	0	1.4	0	22.4	0.2	46.0
Phe2 + Phe3	-	+	+	1.0	1.6	65.5	66.1	99.2	99.2	0	0	0	1.4	0	5.4
	-	+	-	10.6	14.2	98.4	98.8	100	100	0	0.4	0	7.8	0.4	26.0
	-	-	+	4.0	6.4	89.0	88.6	100	100	0	0	0.2	4.0	0	13.0
	-	-	-	0	0	6.6	8.8	59.2	61.8	0	0	0	0	0	0.4

B: Burden test, S: SKAT. Complete: all missing values are filled in with the true values. Incomplete: all missing values are removed.  $\alpha = 1 \times 10^{-5}$ .

Table 4.7: SKAT Power of family data (Imp+Obs) with 20% missing percentage in variant-set tests

Phen Direction	$\varphi_{12}$	$\varphi_{13}$	$\varphi_{23}$	Power (%)					
				100 / 0			50 / 50		
				20%	50%	80%	20%	50%	80%
Complete				77.0	100	100	10.6	50.5	86.4
Incomplete				61.2	100	100	6.9	41.0	76.0
Phe2	+			64.2	100	100	6.7	43.9	77.7
				48.8	99.8	100	2.8	32.9	59.1
Phe3		+		62.2	100	100	6.7	40.2	79.1
				55.7	100	100	2.2	36.2	66.2
Phe2 + Phe3			+	62.9	100	100	7.3	42.6	73.9
				70.5	100	100	5.0	53.8	79.9
				59.7	100	100	3.9	39.8	70.3
				63.4	100	100	3.9	38.4	66.0
				46.0	100	100	2.2	30.4	56.8
				54.8	100	100	3.4	33.7	63.8
				48.3	100	100	2.2	31.5	62.3
				33.5	100	100	0.6	18.8	50.5

Complete: all missing values are filled in with the true values. Incomplete: all missing values are removed.  $\alpha = 1 \times 10^{-5}$ .

Table 4.8: SKAT Power of family data (Imp+Obs) with 50% missing percentage in variant-set tests

Phen Direction	$\varphi_{12}$	$\varphi_{13}$	$\varphi_{23}$	Power (%)					
				100 / 0			50 / 50		
Causal %				20%	50%	80%	20%	50%	80%
Complete				77.7	100	100	7.3	52.1	81.1
Incomplete				32.3	100	100	2.2	20.8	42.6
Phe2	+			45.6	100	100	3.4	27.1	53.3
	-			9.4	93.8	100	0	3.4	16.5
Phe3		+		34.5	100	100	2.6	20.2	42.4
		-		13.7	98.8	100	0.2	7.5	19.4
Phe2 + Phe3	+	+	+	42.4	100	100	3.2	25.3	49.0
	+	+	-	61.0	100	100	2.9	38.6	68.1
	+	-	+	32.4	99.6	100	0.6	16.1	33.3
	+	-	-	27.3	99.8	100	1.4	14.7	38.7
	-	+	+	6.4	87.8	100	0	2.6	11.0
	-	+	-	13.0	98.8	100	0.8	9.7	22.4
	-	-	+	6.6	93.2	100	0.6	2.2	12.4
	-	-	-	5.4	40.4	87.8	0	0.8	1.0

Complete: all missing values are filled in with the true values. Incomplete: all missing values are removed.  $\alpha = 1 \times 10^{-5}$ .

Table 4.9: Variant-set association tests results of FG and 2 hour glucose in FHS

<b>Gene</b>	<b>nsnps</b>	<b>P (FG)</b>	<b>P (2hrglu)</b>
<i>G6PC2</i>	9	0.03	0.93
<i>GCKR</i>	15	0.51	0.42
<i>GLPIR</i>	5	$2.2 \times 10^{-5}$	0.12
<i>VPSI3C</i>	30	0.18	0.21

Table 4.10: Median P-values from FHS 2 hour glucose data in variant-set tests

<b>Dataset</b>	<b>Sample Size</b>	<b><i>G6PC2</i></b>	<b><i>GCKR</i></b>	<b><i>GLPIR</i></b>	<b><i>VPSI3C</i></b>
Obs	5463	0.93	0.42	0.12	0.21
All	5627	0.91	0.41	0.09	0.22
80% Obs + 20% Imp	5463	0.82	0.52	0.08	0.19
80% Obs	4371	0.88	0.42	0.17	0.29
50% Obs + 50% Imp	5463	0.61	0.55	0.03	0.29
50% Obs	2732	0.76	0.61	0.36	0.35

imputation. *GLPIR* has P-values slightly above the  $\alpha$  level of 0.05 in “All”. It shows an improvement in significance compared with “Obs”.

#### 4.4 Discussion

Because rare variants may play an important role in accounting for missing heritability and understanding the biology of the diseases, variant-set association tests have received significant efforts and powerful methods have been developed. However, variant-set tests can still be affected by missingness in the phenotypic data and removing individuals with missing observations can result in a decrease in statistical power. We extended our family-informed phenotype imputation approach derived in Chapter 3 to variant-set association test in genetic studies. The analytical approach to approximate theoretical power for the burden test and SKAT when using the combined observed and imputed

phenotype values was derived. Through extensive simulations, we also evaluated the cases in which our family-informed approach can boost power. Finally, we used a real data example in FHS to illustrate our findings.

By changing the covariance matrix of the phenotype vector and using a non-central Chi-square variable to approximate the distribution of the test statistic, we are able to approximate the power for SKAT theoretically. The simulation results showed that our derivation can provide a very close approximation to the empirical power. We also derived the NCP for burden test which allows different effect sizes for different single variants. Because both the burden test and SKAT statistics can be written as a weighted sum of the test statistics (or squared test statistics) from single-variant test, we found similar conclusions on the performance of our method in variant-set tests. In addition to the correlation between phenotypes and the direction of variant's effect on the phenotypes, proportion of causal variants and proportion of variants with the same direction of effect also affect power in variant-set tests. In the application of FHS 2 hour glucose data, we showed that when the association between the tested gene and the complete phenotype (FG) was strong, our method can help improve significance of the imputed phenotype (2 hour glucose).

Variant-set association tests are more complicated than single variant test because each variant within a gene can act differently, or even have opposite effects. In addition, the

pleiotropy assumption also needs to meet in order to improve statistical significance.

Because SKAT does not yield estimates of effect sizes, we do not need to run the association test on observed data only in order to get an unbiased estimate of effect size as in the single variant test. We chose to use the pooled data of the observed and imputed phenotype values in simulations. Meta-analyzing these two datasets is another option and needs to be examined in the future.

## Chapter 5 Summary and Future Work

In this dissertation, we investigated two issues in genetic association studies: population stratification and missing data. Spurious association results can occur if the population stratification is not corrected. With the increased use of WES genotype data, it is necessary to understand the performance of population stratification adjustment using WES variants. Missing data can cause insufficient statistical power to detect the associations. Even though removing observations with missing values is the most common and easiest way to handle missing data, it could potentially introduce bias when the missing pattern is not missing completely at random (MCAR) and decrease power in association test.

In the first project, two population stratification adjustment methods, PCs and mixed effects models, are examined using GWAS and WES variants, respectively. We found that WES variants have very similar performance as GWAS variants in all evaluations. Hence, computing PCs or GRM with WES variants can capture the population stratification appropriately. When the phenotype is continuous, we observed that LMMs have a higher power than PC-adjusted models for variants confounded by population stratification and these two models perform similarly for variants not confounded by population stratification. When the phenotype is binary, we observed that GLMMs with IBS kinship matrix have inflated type-I error rate using either GWAS or WES variants and hence we suggest to use the BN kinship matrix in GLMMs.

In the second project, we included information contained in family structure and

correlated additional phenotypes in our phenotype imputation method. We derived an approach to approximate the theoretical statistical power of the analysis using the combined observed and imputed data. Based on this derivation of theoretical power and simulations, we identified situations where the statistical significance can be improved after imputation. We also showed that adding information from family structure can increase the imputation accuracy and examined several factors that affect it. In our simulation and real data application, we assumed that the data are missing completely at random. Other missing mechanisms, missing at random (MAR) and missing not at random (MNAR), would require the inclusion of information on other variables and hence need future work.

In the third project, the performance of our family-informed phenotype imputation approach was evaluated in variant-set tests. Approximation to theoretical power were developed for the burden test and SKAT. We examined the performance of our method under different correlation between phenotypes, percentages of missing data, proportion of causal variants and variants with the same direction of effect size. Similar to the second project, we also assumed that the missing mechanism is MCAR. Hence, the performance of our method in MAR and MNAR needs further investigation.

## Bibliography

- [1] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *American Journal of Human Genetics*, 90(1): 7-24, 2012.
- [2] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101(1): 5-22, 2017.
- [3] Bush WS, Moore JH. Genome-Wide Association Studies. *PLoS computational biology*, 8(12): e1002822.
- [4] Patterson NJ, Price AL, Reich D. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006
- [5] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904-909, 2006
- [6] Devlin B, Roeder K. Genomic control for association studies. *Biometrics*, 55(4):997-1004, 1999
- [7] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348-354, 2010
- [8] Conomos MP, Thornton T. GENESIS: GENetic ESTimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. R package version 2.0.1. 2016
- [9] Pritchard J, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*, 155:945-959, 2000
- [10] Belkadi A, Pedergnana V, Cobat A, Itan Y, Vincent QB, Abhyankar A, Shang L, Baghdadi JE, Bousfilha A, the Exome/Array Consortium, Alcais A, Boisson B, Casanova J, Abel L. Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *PNAS*, 113(24):6713-8, 2016
- [11] Gazal S, Gosset S, Verdura E, Bergametti F, Guey S, Babron MC, Tournier-Lasserre E. Can whole-exome sequencing data be used for linkage analysis? *Eur J Hum Genet*, 24(4):581-6, 2016
- [12] Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M. Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol.*, 12(9):R85, 2011
- [13] Kancheva D, Atkinson D, De Rijk P, Zimon M, Chamova T, Mitev V, Yaramis A, Maria Fabrizi G, Topaloglu H, Tournev I, Parman Y, Battaloglu E, Estrada-Cuzcano A, Jordanova A. Novel mutations in genes causing hereditary spastic paraplegia and Charcot-Marie-Tooth neuropathy identified by an optimized protocol for homozygosity mapping based on whole-exome sequencing. *Genet Med.*, 18(6):600-7, 2016

- [14] Eu-ahsunthornwattana J, Miller EN, Fakiola M, Wellcome Trust Case Control Consortium 2, Jeronimo SMB, Blackwell JM, Cordell HJ. Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. *PLoS Genet* 10(7): e1004445, 2014
- [15] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68-74, 2015
- [16] Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304-2305, 2011
- [17] Campbell CD, Ogburn EL, Lunetta KL et al. Demonstrating stratification in a European American population. *Nature Genetics*, 37:868-872, 2005
- [18] Grove ML, Yu B, Cochran BJ, Haritunians T, Bis JC, Taylor KD, Hansen M, Borecki IB, Cupples LA, Fornage M et al. Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium. *PLoS One*, 8(7): e68095, 2013
- [19] International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789-96, 2003.
- [20] Collins AR. Linkage Disequilibrium and Association Mapping: Analysis and Applications. *Human Press*, 2007.
- [21] EPACTS: Efficient and Parallelizable Association Container Toolbox. <http://genome.sph.umich.edu/wiki/EPACTS>
- [22] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. PLINK: a tool set for whole-genome association and population-based linkage-analysis. *American journal of human genetics*, 81(3):559-575, 2007
- [23] Almasy L, Blangero J. Multipoint quantitative trait linkage analysis in general pedigrees. *American journal of human genetics*, 62:1198-122, 1998
- [24] Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*, 11(7): 459463, 2010
- [25] Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*, 46, 100106, 2014
- [26] Liu H, Tang Y, Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics and Data Analysis*, 53:853-856, 2009
- [27] Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89: 82-93, 2011

- [28] Ionita-Laza I, Lee S, Makarov V, Buxbaum J, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*, 92: 841-853, 2013
- [29] Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics*, 95L 5-23, 2014
- [30] Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al.. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91: 224-237, 2012
- [31] Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology*, 37(2): 196-204, 2013
- [32] Chen H, Lumley T, Brody J, et al.. Sequence kernel association test for survival traits. *Genetic Epidemiology*, 38(3): 191-197, 2014
- [33] Lee S, Wu MC, Cai T, Li Y, Boehnke M, Lin X. Power and sample size calculations for designing rare variant sequencing association studies. *Harvard University Technical Report*, 2011
- [34] Rubin DB. Multiple imputation for nonresponse in surveys. *John Wiley & Sons, Inc*, 1987
- [35] Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, Hirschhorn J, Strachan DP, Patterson N, Price AL. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, Volume 30, Issue 20, 15 October 2014, Pages 2906-2914
- [36] Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, Volume 26, Issue 9, 1 May 2010, Pages 1205-1210
- [37] Hormozdiari F, Kang EY, Bilow M, David EB, Vulpe C, McLachlan S, Lusk AJ, Han B, Eskin E. Imputing phenotypes for Genome-wide Association Studies. *American journal of human genetics*, 99: 89-103, 2016
- [38] Dahl A, Iotchkova V, Baud A, Johansson A, Gyllenstein U, Soranzo N, Mott R, Kranis A, Marchini J. A multiple phenotype imputation method for genetic studies. *Nat Genet*, 48(4): 466-472, 2016
- [39] Dupuis J, Siegmund DO, Yakir B. A unified framework for linkage and association analysis of quantitative traits. *PNAS*, 104(51): 20210-20215, 2007
- [40] Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, Lyssenko V, Bouatia-Naji N, Dupuis J, et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet*, 42(2):142-8, 2010

- [41] Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL. Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry*, 63: 1168-174, 2006
- [42] Corder EH, et al.. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimers disease in late onset families. *Science*, 261:921923, 1993
- [43] Genin E, et al. APOE and Alzheimer disease: a major gene with semi-dominant inheritance. *Mol. Psychiatry*, 16:903907, 2011
- [44] Seshadri S, et al.. Genome-wide analysis of genetic loci associated with Alzheimer disease. *J. Am. Med. Assoc*, 303:18321840, 2010
- [45] Naj AC, et al.. Common variants at MS4A4/MS4A6E CD2AP CD33 and EPHA1 are associated with late-onset Alzheimers disease. *Nat Genet*, 43:436441, 2011
- [46] Lambert JC, et al.. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimers disease. *Nat Genet*, 45(12), 14521458, 2013
- [47] Saxena R, et al.. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet*, 42, 142148, 2010
- [48] Wessel J, et al.. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat Commun*, 6: 5897, 2015

