이 학 박 사 학 위 논 문

Statistical Method Development for Rare Variant Association Tests in Family-based Designs

가족 기반 희귀 변이 연관 분석을 위한 분석 알고리즘 개발

2019 년 8 월

서울대학교 대학원

협동과정 생물정보학과

왕 롱 페 이

Statistical Method Development for Rare Variant Association Tests in Family-based Designs

by

Longfei Wang

A thesis
submitted in fulfillment of the requirement
for the degree of Doctor of Philosophy
in
Bioinformatics

Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University
Aug, 2019

Statistical Method Development for Rare Variant Association Tests in Family-based Designs

지도교수 원 성 호

이 논문을 이학박사 학위논문으로 제출함

2019 년 8 월

서울대학교 대학원

생물정보협동과정 생물정보학 전공

왕 롱 페 이

왕롱페이의 이학박사 학위논문을 인준함

2019 년 8 월

위 원 장 박 태 성 (인)

부위원장 원 성 호 (인)

위 원 손 현 석 (인)

위 원 유 연 주 (인)

위 원 박 주 현 (인)

Abstract

Statistical Method Development for Rare Variant Association Tests in Family-based Designs

Longfei Wang

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

Despite of tens of thousands of genome wide association studies (GWASs), the so-called missing heritability reveals that analyses of common variants identified only a limited number of disease susceptibility loci and a substantial amount of causal variants remain undiscovered by GWASs. Sequencing technology was expected to supply this additional information by obtaining large stretches of DNA spanning the entire genome, and improvements in this technology have enabled genetic association analysis of rare/common causal variants. However, single variant association tests commonly used by GWAS result in false negative findings unless very large samples are available. Alternatively, aggregation of association signals across multiple genetic variants in a biology relevant region is expected to boost statistical power for rare variant analysis. Numerous statistical methods have been proposed for

region-based rare variant association studies, such as burden, variance component, and combined omnibus tests.

Region-based association tests are expected to substantially improve statistical power for rare variant analyses and to identify additional disease susceptibility loci. However, very few significant results have been identified due to genetic heterogeneity and relatively small sample sizes. To address the limitations, various approaches have been developed. First, family-based designs play an important role in controlling genetic heterogeneity and population stratification. Second, disease status are often diagnosed by the outcomes of different but related phenotypes, and thus multiple phenotype analysis is supposed to provide additional information and increase power. Third, for the small sample issue, combining results from multiple studies using meta-analysis has been repeatedly addressed as an effective strategy.

In this study, I compared the performance of a selection of the popular family-based rare variant association tests and found *FARVAT* is the most statistically robust and computationally efficient method. Besides, I extended *FARVAT* for multiple phenotype analysis (*mFARVAT*), and meta-analysis (*metaFARVAT*). *mFARVAT* is a quasi-likelihood-based score test for rare variant association analysis with multiple phenotypes, and tests both homogeneous and heterogeneous effects of each variant on multiple phenotypes. *metaFARVAT* combines quasi-likelihood scores from multiple studies and generates burden, variable threshold, variance component, and combined omnibus test statistics. *metaFARVAT* tests homogeneous and heterogeneous genetic effects of variants

among different studies and can be applied to both quantitative and dichotomous phenotypes. With extensive simulation studies under various scenarios, I found that the proposed methods are generally robust and efficient with different underlying genetic architectures, and I identified some promising candidate genes associated with chronic obstructive pulmonary disease, including *DLEC1*.

Key words: rare variant association test, family-based designs, multiple phenotypes, meta-analysis, chronic obstructive pulmonary disease

Student number: 2015-30742

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The background on rare variant association studies

1.1.1 Overview of rare variant association studies

According to genome wide association study (GWAS) Catalog, until 2018 April, 69,885 single nucleotide polymorphism (SNP)-trait associations have been identified by 5,152 GWASs from 3,378 publications. In spite of their success in discovering disease susceptibility loci (DSL), the DSL identified by GWAS have modest effects on disease risk and only partially explain disease heritability. For example, over 70 loci at genome-wide significance only explain 11% of type 2 diabetes heritability (Morris et al. 2012). Rare variants have been implicated as one contributor to this missing heritability (Manolio et al. 2009, Eichler et al. 2010) and have been reported functionally more related

to diseases than common variants (Nejentsev et al. 2009, Price et al. 2010, Genomes Project et al. 2012, Gibson 2012, MacArthur et al. 2012). Recent improvements in DNA sequencing technologies have enabled whole genome sequencing (WGS) studies and more complete assessments of rare genetic variants for modest cost (Cirulli and Goldstein 2010).

However, single variant association analysis under an additive genetic model commonly used by GWAS leads to large false negative findings since the marginal effect of a rare variant cannot be detected unless very large samples are available (Asimit and Zeggini 2010). Moreover, p-values estimates based on regression models might be not accurate if the minor allele count (MAC) is very small (Ma et al. 2013). Alternatively, aggregation of association signals across multiple genetic variants in a biology relevant region, such as a gene, was expected to boost statistical power for rare variant analysis. Numerous methods have been proposed for region-based rare variant association studies and have successfully identified the genetic association of rare variants. These tests can be generally divided into three categories based on the assumptions of the underlying genetic models. The general principles behind these tests are briefly described in Table 1.1.

Most statistical methods for rare variant association tests were propose in a regression framework. Assume *M* markers in a gene are tested with *N* subjects. For subject *i*, the mean of phenotype y_i can be modeled by generalized linear model,

$$g(\mu_i) = \mathbf{X}_i^t \boldsymbol{\alpha} + \mathbf{G}_i^t \boldsymbol{\beta},$$

where $\mathbf{X}_i$ is a vector of covariates, $\mathbf{G}_i$ is allele counts, coded by {0, 1, 2}; $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the regression coefficient vectors for $\mathbf{X}_i$ and $\mathbf{G}_i$, respectively. The score statistic of the marginal model for variant *m* can be defined as

$$S_m = \sum_{i=1}^{N} G_i^m (y_i - \hat{\mu}_i),$$

where $\hat{\mu}_i$ is estimated under the null hypothesis $H_0: \beta_1 = \cdots = \beta_M = 0$. S_m is positive when variant *m* is deleterious and negative when it is protective.

Burden Tests

If we assume that the multiple genetic variants in a region are associated with a trait in the same direction, for instance, all deleterious, we can simply collapse their information into a single genetic score C_i and test the association between C_i and the trait of interest with the simplified model $g(\mu_i) = \boldsymbol{\alpha}^t \mathbf{X}_i + \beta^c C_i$. There are different approaches to define C_i: 1) the cohort allelic sums test (CAST) (Morgenthaler and Thilly 2007) denoted $C_i = 0$ given no minor alleles in a region and $C_i = 1$ otherwise; 2) the combined multivariate and collapsing (CMC) method (Li and Leal 2008) collapsed rare variants into different minor allele frequency (MAF) groups in the same way as CAST; 3) Morris and Zeggini (Morris and Zeggini 2010) assumed a dominant genetic model, in which $C_i = \sum_{m=1}^{M} \mathrm{I}(G_i^m \geq 1)$; 4) the other methods assume an additive genetic model with or without weights in which $C_i = \sum_{m=1}^{M} w_m G_i^m$. Basically, we can define that w_m =1 when MAF less than a fixed threshold and w_m = 0 otherwise. Madsen and Browning (Madsen and Browning 2009) proposed $w_m =$

$1/[p_m(1-p_m)]^{1/2}$, where p_m is the MAF of variant m. Wu et al. (Wu et al. 2011) assumed w_m follows the family of beta densities $w_m = beta(p_m, a_1, a_2)$. In addition, functional effects of variants predicted by bioinformatics tools, such as, PolyPhen (Adzhubei et al. 2010), can be used for weight construction as well. In general, the score statistic to test $H_0: \beta^c = 0$ is defined as

$$Q_{burden} = \left(\sum_{m=1}^{M} w_m S_m \right)^2 \sim \chi^2(df = 1).$$

In addition to the score test, the weighted-sum statistic (WSS) (Madsen and Browning 2009) used the Wilcoxon rank-sum test and calculated p-values by permutation, and the CMC method (Li and Leal 2008) evaluated the joint effect of common variants and rare variant groups using Hoteling's t test. Moreover, the variable threshold (VT) test (Price et al. 2010) applies an optimal frequency threshold instead of a fixed threshold.

Burden tests are powerful when a large fraction of variants are causal and the effects are in the same direction. However, it can lead to a substantial loss of power if this strict assumption is violated. To overcome this limitation, a few two-step methods have been proposed (Han and Pan 2010, Hoffmann et al. 2010, Lin and Tang 2011), which estimate the regression coefficient of each marker first and then assign the weights based on the estimates. Compared to the traditional burden tests, these extensions are more robust, but are unstable for rare variants due to the estimation and computationally expensive because of estimating p-values by permutation.

Variance Component Tests

Variance component (VC) tests, including the sum of squared score (SSU) test (Pan 2009), the sequence kernel association test (SKAT) (Wu et al. 2010), and the C-alpha test (Neale et al. 2011), can also address the limitation of the original burden tests by assuming that the genetic effects in a set follow an arbitrary distribution with mean 0 and variance τ. Therefore, the null hypothesis to be tested becomes $H_0: \tau = 0$, and the VC score statistic is

$$Q_{VC} = \sum_{m=1}^{M} w_m^2 S_m^2 \xrightarrow{d} \sum_{m=1}^{M} \lambda_m \chi_m^2 (df = 1).$$

Q_{VC} asymptotically follows a mixture chi-square distribution with eigenvalues λ_m and calculates p-values analytically. For dichotomous traits, this p-value calculation can produce high false-positive rates if the numbers of cases and controls are imbalance. To overcome this difficulty, Lee et al. proposed a moment-based method that adjusts exact small-sample variance and kurtosis of the test statistic (Lee et al. 2012). VC tests are robust in the presence of both deleterious and protective variants, but less powerful than burden tests when effects are in the same direction.

Combined Omnibus Tests

In practice, the underlying disease architecture is usually unknown. Therefore, it is desirable to propose robust methods for various disease models. A few methods have been proposed to address this difficulty by combining the statistics or the p-values of burden and VC tests. The optimal SKAT (SKAT-O)

(Lee et al. 2012) was proposed as a linear combination of the burden and VC test statistics:

$$Q_{SKATO} = \rho Q_{burden} + (1-\rho)Q_{VC},$$

where $0 \leq \rho \leq 1$ can be interpreted as a pairwise correlation among β_m and estimated by minimizing p-value with a grid of ρs. The asymptotic p-value of SKAT-O can be calculated with computationally efficient one-dimensional numerical integration.

Another approach is to combine the p-values of the two tests using Fisher's method and calculate its p-value by permutation (Derkach et al. 2013). The Fisher statistic is

$$\text{Fisher} = 2\log(P_{burden}) - 2\log(P_{VC}),$$

where P_{VC} and P_{burden} are the p-values calculated from burden and VC tests, respectively. To reduce the computational intensive, Sun et al. (Sun et al. 2013) derived the asymptotic p-value by modifying VC statistic to make it independent of burden test statistic.

Combined omnibus tests are more robust with respect to the unknown disease architecture, but can be slightly less powerful than burden or VC tests if their assumptions are largely held.

Table 1.1 Rare variant association test methods.

Category	Method	Reference
Burden test	CAST: cohort allelic sums test	(Morgenthaler and Thilly 2007)
	CMC: combined multivariate & collapsing	(Li and Leal 2008)
	WSS: weighted-sum statistic	(Madsen and Browning 2009)
	MZ: Morris and Zeggini	(Morris and Zeggini 2010)
	VT: variable threshold	(Price et al. 2010)
	aSum: data-adaptive sum test	(Han and Pan 2010)
	Step-up: model-selection framework	(Hoffmann et al. 2010)
	EREC: estimated regression coefficient	(Lin and Tang 2011)
Variance component tests	SSU: sum of squared score	(Pan 2009)
	SKAT: sequence kernel association test	(Wu et al. 2010)
	C-alpha: C-alpha score test	(Neale et al. 2011)
Combined omnibus tests	SKAT-O: optimal SKAT	(Lee et al. 2012)
	Fisher's method	(Derkach et al. 2013)
	MiST: mixed-effects score test	(Sun et al. 2013)

1.1.2 Challenges of rare variant association studies

Aggregation of association signals across multiple genetic variants is expected to substantially increase statistical power for rare variant analysis and to identify additional DSL. However, the rare variant association tests with population-based samples suffer from genetic heterogeneity due to population substructure and admixture. Moreover, it can lead to loss of power when a very few variants in a region are associated with the trait of interest and result in inaccurate type I error (TIE) rates when MACs are very small. Therefore, the approaches to control genetic heterogeneity and enrich genetic effects are desirable. Here, I discuss three approaches: family-based designs, multiple phenotype analysis, and meta-analysis.

Family-based designs

Various study designs have been developed to minimize genetic heterogeneity, such as selecting individuals with extreme phenotypes, (Merikangas et al. 1989, Goldin et al. 1991). In families, Mendelian transmission results in family members sharing the same alleles, and thus, affected family members have a greater chance to carry the same causal variants than unrelated subjects (Shi and Rao 2011). Therefore, genetic heterogeneity among affected relatives is expected to be smaller, and family-based designs have been repeatedly addressed as an important strategy for rare variant association studies. Numerous family-based methods have been proposed, such as, the transmission disequilibrium test (TDT) method (Spielman et al. 1993), generalized estimating equations (GEE) (Chen and Yang 2010) and mixed

models (Slager and Schaid 2001, Bourgain et al. 2003, Thornton and McPeek 2007, Choi et al. 2009), which will be reviewed in the next chapter.

Multiple phenotype analysis

Genetic association analyses simultaneously test a large number of variants, and stringent significance levels imposed by the multiple testing problem highlight the importance of powerful strategies. Multiple measurements can be obtained from different but related phenotypes, or from repeated measurements of a single phenotype at different time points. In particular, disease diagnose is usually based on a number of different phenotypes. Association analyses with multiple phenotypes often lead to substantial improvements in statistical power (Schifano et al. 2013) and such improvements are inversely related to correlations between phenotypes (Lee et al. 2014). A few different methods have been proposed, including the scaled marginal model (Schifano et al. 2013) and the extended Simes procedures for population-based samples (van der Sluis et al. 2013). The statistical power of these methods depends on the underlying genetic architectures between the causal variants and the multiple phenotypes, either homogeneous or heterogeneous (van der Sluis et al. 2013). Won et al. proposed an omnibus family-based association test for the joint analysis of multiple genotypes and multiple phenotypes (MFQLS) for common variant analysis (Won et al. 2015) and identified intronic variant pair on SIDT2 associated with metabolic syndrome in a Korean population (Moon et al. 2018). However, a very few multiple phenotype analyses have been developed for rare variant studies.

Meta-analysis

When the sample sizes are small, statistical analyses suffer from high false negative error rates, and this limitation can be avoided by combining data from multiple studies via mega- or meta-analysis. Mega-analysis assumes that subjects' genotypes and phenotypes from different studies are available, and these are pooled for genetic association analyses. Meta-analysis directly utilizes test statistics from separate studies and combines them into a single test statistic. The choice between mega- and meta-analysis depends on the heterogeneity among studies and the availability of individual genotype and phenotype data from all studies. Particularly, if there are systematic differences in phenotype diagnosis or sequencing platforms, meta-analysis is often preferred. Furthermore, it has been proved meta-analysis can be as powerful as mega-analysis (Lee et al. 2013, Liu et al. 2014). Recently, several meta-analysis methods for rare variant association tests have been proposed (Table 1.4), such as MASS (Tang and Lin 2013, Tang and Lin 2014), RAREMETAL (Feng et al. 2014), seqMeta (Chen et al. 2014), and metaSKAT (Lee et al. 2013). However, the available statistical methods for family-based samples or dichotomous phenotypes are limited, and thus, it is worthwhile to provide a method that can be applied to both quantitative and dichotomous phenotypes under homogeneous (hom) and heterogeneous (het) genetic effect models.

Table 1.2 Meta-analysis for rare variant association tests

Method	Phenotype		Study design		Test		Reference
	quantitative	dichotomous	unrelated	families	homogenous	heterogeneous	
MASS	√	√	√		√	√	(Tang and Lin 2013)
metaSKAT	√	√	√	quantitative	√	√	(Lee et al. 2013)
seqMeta	√	√	√		√		(Chen et al. 2014)
RAREMETAL	√		√	quantitative	√		(Feng et al. 2014)

The definition of the acronyms in Table 1.2: 1) MASS: the meta-analysis of score statistics for sequencing studies; 2) metaSKAT: the meta-analysis for SNP-set (sequence) kernel association test; 3) seqMeta: the meta-analysis of region-based tests of rare DNA variants; 4) RAREMETAL: a tool for meta-analysis of rare variants using sequencing or genotyping array data.

1.2 Purpose of this study

The main purpose of this thesis is to develop statistical methods for detecting rare variant associations and enriching genetic effects by using family-based designs, multiple phenotypes, and meta-analysis. In this thesis, first I compared the existing family-based rare variant association studies and found *FARVAT* (Choi et al. 2014) is the most powerful, robust, computationally efficient method.

Second, I proposed a multivariate family-based rare variant association tool (*mFARVAT*). Human diseases are often defined by the outcomes of multiple phenotypes, and thus I expect multivariate family-based analyses may be very efficient in detecting associations with rare variants. However, few statistical methods implementing this strategy have been developed for family-based designs. Therefore, I proposed the *mFARVAT,* which is a quasi-likelihood-based score test for rare variant association analysis with multiple phenotypes, and tests both homogeneous and heterogeneous effects of each variant on multiple phenotypes. Simulation results show that the proposed method is generally robust and efficient for various disease models, and I identify some promising candidate genes associated with chronic obstructive pulmonary disease (COPD).

Third, I proposed a family-based rare variant association test for meta-analysis (*metaFARVAT*). Although, family-based designs have been shown to be powerful in detecting the significant rare variants associated with human

diseases, very few significant results have been found owing to relatively small sample sizes and the fact that statistical analyses often suffer from high false-negative error rates. These limitations can be overcome by combining results from multiple studies via meta-analysis. However, statistical methods for meta-analysis with rare variants are limited for family-based samples. Therefore, I proposed *metaFARVAT*. By combining the scores calculated from each study using *FARVAT*, *metaFARVAT* generates burden test, VT test, SKAT, and SKAT-O statistics. The proposed method tests homogeneous and heterogeneous effects of variants among different studies and can be applied to both quantitative and dichotomous phenotypes. Simulation results demonstrated the robustness and efficiency of the proposed method in different scenarios. By applying *metaFARVAT* to data from a family-based study and a case-control study, I identified a few promising candidate genes, including *DLEC1*, which is associated with COPD.

Last, both of the proposed methods were applied to chronic obstructive pulmonary disease (COPD) data. COPD is a type of obstructive lung disease characterized by long-term breathing problems and poor airflow. There are two main measurements for diagnosis, the forced expiratory volume in one second (FEV1), which is the greatest volume of air that can be breathed out in the first second of a breath, and the forced vital capacity (FVC), which is the greatest volume of air that can be breathed out in a single large breath. Normally, 75–80% of the FVC comes out in the first second and a FEV1/FVC ratio <70% in someone with symptoms of COPD defines a person as having the disease. The

Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines suggest dividing people into four categories based on symptoms assessment and airflow limitation: 1) Mild: GOLD = 1 if FEV1 ≥80%; 2) Moderate: GOLD = 2 if 50% ≤ FEV1 ≤ 79%; 3) Servere: GOLD = 3 if 30% ≤ FEV1 ≤ 49%; 4) Very severe: GOLD = 4 if FEV1 < 30%. As of 2018, COPD affected about 328 million of the global population. In 2018 only, it resulted in about 4 million deaths. In the United States, COPD is estimated to be the third leading cause of death, approximately 6.3% of the adult population, totaling approximately 15 million people, have been diagnosed with COPD. Smoking is the main risk factor of COPD. Genetics play a role in the development of COPD. Alpha 1-antitrypsin deficiency has been proven as a genetic factor. The disease risk is particularly high if someone who is deficient in alpha 1-antitrypsin also smokes. Therefore, it is worth to investigate other possible genetic factors.

1.3 Outline of the thesis

This thesis is organized as follows. Chapter 1 is an introduction to this study with an overview of the existing rare variant association studies, the challenges and the approaches to enrich genetic effects. Chapter 2 consist of an overview of family-based association studies and a comparison of the existing family-based rare variant association methods with GAW19 data. Chapter 3 is an extension of *FARVAT* for multiple phenotype analysis. Chapter 4 is an extension for meta-analysis. Chapters 3 and 4 contain introductions to the statistical methods, simulation studies, and the applications to COPD data. Finally, the summary and conclusions are presented in Chapter 5.

This chapter was published in *BMC proceedings*
as a partial fulfillment of Longfei Wang's PhD program.

Chapter 2

Overview of family-based rare variant association tests

2.1 Overview of family-based association studies

Family-based design is commonly used in genetic association studies. The current statistical methods for family samples can be grouped into two major categories referred to as conditional methods and unconditional methods.

Conditional methods

The conditional family-based design is based on evaluating the association between a phenotype and the transmission of marker alleles within family members. The popular methods for single SNP analysis are the transmission disequilibrium test (TDT) method (Spielman et al. 1993) and its extensions,

such as family-based association test (FBAT) (Laird and Lange 2006). These tests compare the observed number of alleles of type 1 that are transmitted to the affected offspring with those expected from Mendelian transmissions. An excess of type 1 alleles among the affected indicates that a DSL for the trait is linked and associated with the marker locus. Therefore, the test statistics model the offspring genotypes conditional on informative/heterozygous parental genotypes within each trio, and preserve inherent robustness against population heterogeneity. TDT has been extended for rare variant analysis (Derkacheva and Hennig 2014). FBAT statistics also have been extended for joint analysis of multivariate phenotypes and genotypes (Gray-McGuire et al. 2009), and for rare variants (Yip et al. 2011). However, they do not fully use the information in the parental phenotypes, and loss of power can be substantial if the number of founders is relatively large.

Unconditional methods

Unconditional methods directly model the associations between phenotypes and genotypes of all individuals and incorporate both population and pedigree structure using a covariance matrix, which can be constructed with known structure or estimated from genome-screen data. The pedigree information is defined as the kinship matrix. For instance, in family *i*,

$$\mathbf{\Phi}_i = \begin{bmatrix} 1 + h_{i1} & \cdots & 2\phi_{i1n} \\ \vdots & \ddots & \vdots \\ 2\phi_{in1} & \cdots & 1 + h_{in} \end{bmatrix},$$

where h_{ij} is the inbreeding coefficient of individual *j* in family *i*, and ϕ_{ijk} is the kinship coefficient between individuals *j* and *k* in family *i*.

The correlation among family members can be taken into account in GEE (Chen and Yang 2010). μ_{ij}, the mean of the phenotype y_{ij} of individual *j* in family *i*, was modeled using the marginal generalized linear model

$$g(\mu_{ij}) = \mathbf{X}_{ij}^t\boldsymbol{\alpha} + \mathbf{G}_{ij}^t\boldsymbol{\beta}.$$

The link function $g(\cdot)$ is μ_{ij} for continuous phenotype and is $\text{logit}(\mu_{ij})$ for dichotomous phenotypes. The GEE for the parameters can be written as

$$\text{U}(\boldsymbol{\theta}) = \text{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i^t \mathbf{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \sum_{i=1}^{n} \begin{pmatrix}\mathbf{X}_i \\ \mathbf{G}_i\end{pmatrix}^t \boldsymbol{\Delta}_i \mathbf{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i),$$

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\theta}^t$; $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\delta)\mathbf{A}_i^{1/2}$ is a working covariance matrix of y_i; $\text{A}_i = \text{diag}\{v(\mu_{i1}), \ldots, v(\mu_{im})\}$, where $v(\mu_{ij})$ is a variance function. $\mathbf{R}_i(\delta)$ is a working correlation matrix with the theoretical kinship coefficient $\boldsymbol{\Phi}_i$ and a scale parameter δ. $\boldsymbol{\Delta}_i = \text{diag}\{\dot{\mu}_{i1}, \ldots, \dot{\mu}_{im}\}$, where $\dot{\mu}$ is the first derivative of $g^{-1}(\cdot)$. The GEE method was extended for rare variant analysis with variance component test (Wang et al. 2013).

The unbalanced nature of family-based samples can lead to bias of sandwich estimators for the variance-covariance matrix, and results from GEE can be invalid (Aaij et al. 2013). An alternative approach is to take the covariance matrix into a generalized linear mixed model (GLMM) framework by including a random polygenic effect ***b***,

$$g(\mu_{ij}) = \mathbf{X}_{ij}^t\boldsymbol{\alpha} + \mathbf{G}_{ij}^t\boldsymbol{\beta} + \boldsymbol{b}_{ij}.$$

where $\boldsymbol{b} \sim MVN(0, \sigma^2\boldsymbol{\Psi})$, where σ is the variance of $\mathbf{G}$ for an outbred individual in the absence of population structure and $\boldsymbol{\Psi}$ accounts for relatedness, inbreeding, and population structure. When structure is known, the estimator of σ^2 is $\hat{\sigma}_1{}^2 = (n-1)^{-1}[\mathbf{G}^t\boldsymbol{\Psi}^{-1}\mathbf{G} - (1^t\boldsymbol{\Psi}^{-1}1)^{-1}(1^t\boldsymbol{\Psi}^{-1}\mathbf{G})^2]$ or $\hat{\sigma}_2{}^2 = 2\hat{p}(1-\hat{p})$ if Hardy-Weinberg equilibrium (HWE) holds at the marker, where $\hat{p}$ is a suitable estimator of MAF, such as the sample frequency $\hat{p} = \overline{\mathbf{G}}$, or the best linear unbiased estimator $\hat{p} = (1^t\boldsymbol{\Psi}^{-1}1)^{-1}1^t\boldsymbol{\Psi}^{-1}\mathbf{G}$.

Several methods have been propose for association testing in related samples with the assumption of no additional population structure, which is $\boldsymbol{\Psi} = \boldsymbol{\Phi}$, including the corrected Pearson χ^2 test, the Armitage trend test, the W_{QLS} test, and the M_{QLS} test. (Slager and Schaid 2001, Bourgain et al. 2003, Thornton and McPeek 2007, Choi et al. 2009) and their test statistics can be generalized as

$$\frac{(\mathbf{T}^t\mathbf{G})^2}{(\hat{\sigma}^2\mathbf{T}^t\boldsymbol{\Phi}\mathbf{T})} \sim \chi^2(df = 1),$$

where $\mathbf{T}$ is a fixed, nonzero column vector including phenotype information, or additionally including pedigree or covariate information or both. Specifically, 1) in the corrected Pearson χ^2 and the Armitage trend test: $\mathbf{T} = \mathbf{1}_c - \frac{n_c}{n}\mathbf{1}$, where $\mathbf{1}_c$ is 1 if individual *i* is a case and 0 if a control, n_c is the number of cases; 2) in W_{QLS} test: $\mathbf{T} = \boldsymbol{\Phi}^{-1}\mathbf{1}_c - \mathbf{1}_c^t\boldsymbol{\Phi}^{-1}\mathbf{1}(\mathbf{1}^t\boldsymbol{\Phi}^{-1}\mathbf{1})^{-1}\boldsymbol{\Phi}^{-1}\mathbf{1}$; 3) in M_{QLS} test: $\mathbf{T} = \mathbf{A}_N + \boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}_{N,M}\mathbf{A}_M - \left(\mathbf{A}_N + \boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}_{N,M}\mathbf{A}_M\right)^t\mathbf{1}(\mathbf{1}^t\boldsymbol{\Phi}^{-1}\mathbf{1})^{-1}\boldsymbol{\Phi}^{-1}\mathbf{1}$,

where $\mathbf{A}_N$ is the phenotype vector with non-missing genotype and $\mathbf{A}_M$ is the one with missing genotype; the phenotype vector is coded as $A_i = 1$ if *i* is affected, $A_i = -k/(1-k)$ if *i* is unaffected, where *k* is prevalence; $A_i = 0$ if *i* is missing. However, these tests tend to have inflated TIE in the presence of population heterogeneity. Thornton and McPeek proposed ROADTRIPS (Thornton and McPeek 2010) to extend the above tests with the estimator $\hat{\mathbf{\Psi}}$ from genome-screen data to simultaneously correct for both population and pedigree structure,

$$\hat{\mathbf{\Psi}}_{ij} = \begin{cases} \frac{1}{M}\sum_{m=1}^{M} \frac{(G_i^m - 2\hat{p}_m)(G_j^m - 2\hat{p}_m)}{2\hat{p}_m(1-\hat{p}_m)}, i \neq j \\ 1 + \frac{1}{M}\sum_{m=1}^{M} \frac{{G_i^m}^2 - (1+2\hat{p}_m)G_i^m + 2\hat{p}_m^2}{2\hat{p}_m(1-\hat{p}_m)}, i = j \end{cases}$$

where $\hat{p}_m = \bar{\mathbf{G}}^m$. Accordingly, the estimator of σ^2 becomes ${\hat{\sigma}_1}^2 = (n-1)^{-1}\mathbf{G}^t\hat{\mathbf{\Psi}}^{-}\mathbf{G}$, where $\hat{\mathbf{\Psi}}^{-}$ is the Moore-Penrose generalized inverse of $\hat{\Psi}$. An alternative way to estimate $\hat{\mathbf{\Psi}}_{ij}$ is based on estimated probabilities of identical by descent (IBD) sharing using moment-based (Purcell et al. 2007) or maximum likelihood estimation (Sun et al. 2002, Weir et al. 2006). The mixed model methods have gained increasing popularity recently because they are computationally efficient and easy to integrate data with both family and unrelated individuals. Numerous methods have been proposed for rare variant association tests based on this mixed model framework (Schaid et al. 2013, Choi et al. 2014).

2.2 Comparison of the selected family-based rare variant association tests

I selected a number of family-based rare variant association methods from different categories (Table 2.2) and compared their performance for dichotomous phenotype analysis using Genetic Analysis Workshop 19 (GAW19) simulated data. I considered five different methods: the Rare Variant Transmission Disequilibrium Test (RV-TDT) (Derkacheva and Hennig 2014), the GEE-based Kernel Association (GEE-KM) test (Wang et al. 2013), an extended Combined Multivariate and Collapsing test for Pedigrees (PedCMC) (Zhu and Xiong 2012), the Gene-level kernel and burden tests for Pedigrees (PedGene) (Schaid et al. 2013), and the FAmily-based Rare Variant Association Test (*FARVAT*) (Choi et al. 2014). These methods were utilized to identify causal genes for hypertension, and the results were compared in regard to their statistical and computational efficiency. Our results showed that PedGene and *FARVAT* are usually the most statistically powerful, and with regards to the computational intensity, *FARVAT* is the most efficient.

Table 2.1 Family-based common variant association tests.

Category		Method	Reference
Conditional	TDT	TDT: transmission disequilibrium test	(Spielman et al. 1993)
		FBAT: family-based association test	(Laird and Lange 2006)
Unconditional	GEE	GEE: generalized estimating equations	(Chen and Yang 2010)
	GLMM	Armitage trend test	(Sasieni 1997)
		corrected Pearson χ^2 test	(Slager and Schaid 2001)
		W_{QLS} test: quasi-likelihood score test	(Bourgain et al. 2003)
		M_{QLS} test: quasi-likelihood score test	(Thornton and McPeek 2007)
		ROADTRIPS: robust association-detection test for related individuals with population structure	(Thornton and McPeek 2010)

Table 2.2 Family-based rare variant association tests.

Category		Method	Reference
Conditional	TDT	RV-TDT: Rare Variant Transmission Disequilibrium Test	(Derkacheva and Hennig 2014)
		RVGDT: Rare Variant Generalized Disequilibrium Test	(He et al. 2017)
Unconditional	GEE	GEE-KM: Generalized Estimating Equations based Kernel Machine test	(Wang et al. 2013)
	GLMM	PedCMC: Combined Multivariate and Collapsing test for Pedigrees	(Zhu and Xiong 2012)
		PedGene: Gene-level kernel and burden tests for Pedigrees	(Schaid et al. 2013)
		FARVAT : FAmily-based Rare Variant Association Test	(Choi et al. 2014)
		FSKAT: Sequence Kernel Association Tests for families	(Yan et al. 2015)

2.2.1 Rare Variant Transmission Disequilibrium Test (RV-TDT)

RV-TDT (Derkacheva and Hennig 2014) is an extension of TDT (Spielman et al. 1993) to analyze parent-child trio data for rare variant associations, which can adequately control for population admixture. RV-TDT is implemented with C and can calculate four burden test methods: CMC, WSS, burden of rare variants (BRV), and VT.

For parent *i* with variant *m*, the indicator variables $c_i^m = 1$ if a minor-allele-transmitted event occurs, and $b_i^m = 1$ if a major-allele-transmitted event occurs and otherwise 0. For a genetic region *L*, the total minor-allele-transmitted events and major-allele-transmitted events for parent *i* are given by

$$c_i = \sum_{m \in L} c_i^m \, , b_i = \sum_{m \in L} b_i^m .$$

With *n* trios, for the TDT-CMC test, each informative parent contributes a score of 1 to the McNemar's test. The statistics are given by

$$c = \sum_{i=1}^{2n} c_i/(b_i + c_i) \, , b = \sum_{i=1}^{2n} b_i/(b_i + c_i).$$

For the TDT-BRV method, each informative parent contributes a score that equals to the number of informative sites within the region and thus *c* and *b* are given in the form of

$$c = \sum_{i=1}^{2n} c_i \, , b = \sum_{i=1}^{2n} b_i .$$

For the TDT-WSS, each variant site is weighted by $\widehat{w}_m = \sqrt{n_m q_m (1 - q_m)}$, where q_m is the allele frequency of variant m in parental haplotypes that are not transmitted to the offspring. c and b are given by

$$c = \sum_{i=1}^{2n} \sum_{m \in L} \frac{c_i^m}{\widehat{w}_m}, b = \sum_{i=1}^{2n} \sum_{m \in L} \frac{b_i^m}{\widehat{w}_m}.$$

For the TDT-VT test, the test statistic is maximized over allele frequencies and therefore, a variable allele frequency threshold is applied, instead of a fixed MAF cut-off.

2.2.2 Generalized Estimating Equations based Kernel Machine test (GEE-KM)

Wang et al. (Wang et al. 2013) proposed a family-based kernel machine (KM) (Wu et al. 2010) SNP set test in the GEE framework for both continuous and dichotomous phenotypes. In addition, Wang et al. developed analytical methods to calculate the p-values and proposed a resampling method for correcting for small sample size bias in family studies. GEE-KM can adjust for the effect of covariates aFnd was implemented in the gskat R package.

With the assumption that $\beta_m \, (m = 1, \dots, M)$ follow an arbitrary distribution with mean 0 and common variance τ, the null hypothesis is $H_0 : \tau = 0$. Therefore, based on the GEE framework introduced in the previous section, the KM test is

$$Q_S = \widetilde{\mathbf{U}}_G^t \widetilde{\mathbf{U}}_G \xrightarrow{d} \sum_{m=1}^{M} \lambda_m \chi_{m,1}^2 ,$$

where $\widetilde{\mathbf{U}}_G = \sum_{i=1}^{n} \mathbf{G}_i^t \Delta_i \mathbf{V}_i^{-1} (\boldsymbol{y}_i - \widetilde{\boldsymbol{\mu}}_i)$, where $\widetilde{\boldsymbol{\mu}}_i = g^{-1}(\mathbf{X}_i^t \widetilde{\boldsymbol{\alpha}})$; $\chi_{m,1}^2$ are independent $\chi^2(df = 1)$ random variables; and λ_m are eigenvalues. The p-value adjusted for small samples can be calculated as

$$1 - F\left((Q_S - \hat{\mu}_Q)\sqrt{2df}/\sqrt{\hat{v}_Q} + df \,|\chi_{df}^2\right),$$

where $F((\cdot\,|\chi_{df}^2)$ is the distribution of χ_{df}^2 and $df = 12/\hat{\gamma}$; $\hat{\mu}_T$, $\hat{v}_Q$ and $\hat{\gamma}$ are the estimated small sample mean, variance and kurtosis of the statistic Q_S under the null, respectively.

2.2.3 Combined Multivariate and Collapsing test for Pedigrees (PedCMC)

PedCMC (Zhu and Xiong 2012) was proposed as an combination of the collapsing test (Li and Leal 2008) and the population-based generalized T^2 test (Xiong et al. 2002) for pedigrees. The genotypes for rare variants in each gene are coded as either 0 or 1 according to the presence of rare alleles, and the sums of coded genotypes are compared between affected and unaffected individuals.

The indicator variable $v_i^s = 1$ if rare variants in group *s* of individual *i* is present. *M* variants G are consist of *k* groups of rare variants **X** and *m* individual variant sites **Z**, $\mathbf{G} = [\mathbf{X}, \mathbf{Z}]^t$.

$$\mathbf{H}_{CMC} = \mathbf{I}_{(k+m)} \otimes \mathbf{T}^t,$$

where $\mathbf{T} = \mathbf{1}_c - \frac{n_c}{n}\mathbf{1}$ and $\otimes$ denotes the Kronecker product.

$$\mathbf{\Gamma}_{CMC} = \mathbf{T}^t\mathbf{\Phi}\mathbf{T}\mathbf{\Sigma},$$

where $\mathbf{\Phi}$ is the estimated kinship matrix proposed by Thornton (Thornton and McPeek 2010) and $\mathbf{\Sigma}$ is the covariance matrix of genotypes, which is define as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_{xz} \\ \mathbf{\Sigma}_{zx} & \mathbf{\Sigma}_z \end{bmatrix}.$$

Therefore, the family-based CMC statistic can be defined as

$$T^2_{CMCF} = (\mathbf{H}_{CMC}\mathbf{G})^t\mathbf{\Gamma}^{-1}_{CMC}(\mathbf{H}_{CMC}\mathbf{G})$$

$$= \frac{\frac{n_c(n-n_c)}{n}[(\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_B)^t\mathbf{\Sigma}_v^{-1}(\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_B) + (\bar{\mathbf{Z}}_A - \bar{\mathbf{Z}}_B)^t\mathbf{\Sigma}_z^{-1}(\bar{\mathbf{Z}}_A - \bar{\mathbf{Z}}_B)]}{\frac{n}{n_c(n-n_c)}\mathbf{T}^t\mathbf{\Phi}\mathbf{T}}$$

$$= \frac{T^2_{CMC}}{P_{corr}} \sim \chi^2(df = k + m).$$

where $\bar{\mathbf{X}}_A, \bar{\mathbf{X}}_B$ are the average of the indicator variables for the rare variant groups in cases and controls, respectively; $\bar{\mathbf{Z}}_A, \bar{\mathbf{Z}}_B$ are the average of the indicator variables for the genotypes in cases and controls, respectively; T^2_{CMC} is the CMC statistic for the population-based association test; and P_{corr} is the correlation factor to be applied to the generalized T^2 statistic to have a valid test in the presence of pedigree structures.

2.2.4 Gene-level kernel and burden tests for Pedigrees (PedGene)

Schaid D.J. et al (Schaid et al. 2013) proposed burden and kernel statistics for extended families, and it was implemented in the PedGene R package. This

approach views the sample collection as a retrospective study, which means conditioning on phenotypes and treating the genotype data random.

The covariance of the genotype for subject *j* and *j'*, and markers *m* and *m'*, can be expressed as

$$Cov_o(g_{j,m}, g_{j',m'}) = \sigma_{mm'}\mathbf{\Phi}_{jj'} = 2\rho_{mm'}\sqrt{p_m(1-p_m)p_{m'}(1-p_{m'})}\mathbf{\Phi}_{jj'},$$

where *p* is the MAF for the markers, ρ is the correlation of genotype. $\mathbf{\Phi}$ is the estimated kinship matrix proposed by Thornton (Thornton and McPeek 2010). The kernel statistic is

$$Q_{kernel} = \left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)^t \mathbf{GWWG}^t\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right),$$

where $\mathbf{Y}$ is a vector of disease status indicators of *n* subjects; $\mathbf{Y} - \widehat{\mathbf{Y}}$ is the vector of residuals after adjusting for covariates by use of logistic regression models; and $\mathbf{W}$ is a diagonal matrix with weights for each markers. The distribution of Q_{kernel} was estimated by a scaled distribution with the scale $\delta = Var[Q_{kernel}]/(2E[Q_{kernel}])$ and the $df = 2E[Q_{kernel}]^2/Var[Q_{kernel}]$. P-values were computed by assuming $Q_{scaled} = Q_{kernel}/\delta \sim \chi^2_{df}$.

The statistic for burden test is

$$Q_{burden} == \frac{\left[\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)^t \mathbf{GW1}_M\right]^2}{(\mathbf{1}_M^t \mathbf{W\Sigma W1}_M)\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)^t \mathbf{\Phi}\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)} \sim \chi^2(df = 1).$$

2.2.5 FAmily-based Rare Variant Association Test (*FARVAT*)

FARVAT (Choi et al. 2014) is a family-based rare variant association test based on the quasi-likelihood (Thornton and McPeek 2007) and provides burden, C-alpha and SKAT-O type statistics (Lee et al. 2012) for quantitative and dichotomous phenotypes. *FARVAT* was implemented with C++.

The score for the quasi-likelihood (Thornton and McPeek 2007) is

$$\mathbf{T}^t\mathbf{V}^{-1}\left(\mathbf{G}-\hat{E}(\mathbf{G})\right)=\mathbf{T}^t(\mathbf{I}_N-\mathbf{1}_N(\mathbf{1}_N^t\mathbf{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\mathbf{\Phi}^{-1})\mathbf{G}.$$

where **T** is the phenotype adjusted by offset which can be prevalence, which is equivalent to $\mathbf{T}=\mathbf{1}_c-\frac{N_c}{N}\mathbf{1}$, or best linear unbiased prediction (BLUP) $\mathbf{T}=\mathbf{Y}-\widehat{\mathbf{Y}}$; Φ is the estimated kinship matrix proposed by Thornton (Thornton and McPeek 2010). Therefore, I have

$$\frac{1}{\sqrt{\mathbf{T}^t\mathbf{A}\mathbf{T}}}\mathbf{T}^t\mathbf{A}\mathbf{\Phi}^{-1}\mathbf{G}\mathbf{\Sigma}^{1/2}\sim MVN(0,\mathbf{I}_M)\text{ under }H_0.$$

where $\mathbf{A}=\mathbf{\Phi}-\mathbf{1}_N(\mathbf{1}_N^t\mathbf{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t$ and $\mathbf{\Sigma}$ is the covariance matrix of genotypes. If the weight for variant *m* is denoted as w_m, the null hypotheses for the burden test and the C-alpha test are respectively:

$$H_0^1: w_1\beta_1+\cdots+w_M\beta_M=0,$$

$$H_0^2: w_1^2\beta_1^2+\cdots+w_M^2\beta_M^2=0.$$

Therefore, both score tests for rare variant analysis can be generalized to

$$S_\rho=\frac{1}{\mathbf{T}^t\mathbf{A}\mathbf{T}}\mathbf{T}^t\left(\mathbf{G}-\hat{E}(\mathbf{G})\right)\mathbf{W}[(1-\rho)\mathbf{I}_M+\rho\mathbf{1}_M\mathbf{1}_M^t]\mathbf{W}\left(\mathbf{G}-\hat{E}(\mathbf{G})\right)^t\mathbf{T},$$

where $\rho \in [0,1]$, when $\rho = 1$, S_1 is the score for burden test, while $\rho = 0$, S_0 is the score for C-alpha test. The eigenvalues for $\Sigma^{1/2}WW\Sigma^{1/2}$ is written as λ_m. Therefore,

$$S_1 \sim (\mathbf{1}_M^t \mathbf{W\Sigma W 1}_M)\chi^2(df = 1) \text{ under } H_0^1,$$

$$S_0 \sim \sum_{m=1}^{M} \lambda_m \chi_m^2(df = 1) \text{ under } H_0^2.$$

For $\rho_0 = 0 < \rho_1 < \cdots < \rho_L = 1$, The observed value for S_{ρ_l} is denoted by s_{ρ_l}, and their corresponding *p*-values are denoted by p_{ρ_l}. Furthermore, the (1 – *p*)*th* quantile for S_{ρ_l} is written as $Q_{\rho_l}(p)$. Therefore, the SKAT-O statistic (Lee et al. 2012) is

$$p_{min} = \min\{p_{\rho_0}, p_{\rho_l}, \dots, p_{\rho_L}\},$$

and its *p*-value is obtained by

$$1 - P\left(S_{\rho_0} < Q_{\rho_0}(p_{min}), \dots, S_{\rho_L} < Q_{\rho_L}(p_{min})\right).$$

2.2.6 Comparison of the methods with GAW19 data

To access performance of methods, a simulated data set of 200 phenotype replicates was provided for the family data sets. It was based on the real data, with the family structure, sex, and age taken from the real data. Blood pressure, medication use, and tobacco smoking were generated for each replicate, using the distributional structure found in the real data. The simulated values of systolic (SBP) and diastolic blood pressure (DBP) were influenced by over

1000 variants in over 200 genes. Individuals with SBP<140 or DBP>90 were assigned to be affected by hypertension. In addition, a normally distributed trait, Q1, was simulated that was not influenced by any genetic variants, but was correlated between family members (Engelman et al. 2016). Genotypes for 959 individuals imputed from 464 sequenced subjects were used in our analysis, and I considered rare variants of which MAF <0.05. Rare variants were annotated with High, Moderate, and Low risk effect by using SnfEff software (Cingolani et al. 2012), and those variants were used for gene-set analysis. The set file included 58,969 SNPs in 7,210 genes, which was used to evaluate the statistical validity for all the methods.

For the evaluation of statistical validity, the empirical TIE estimates for all the methods were calculated at various significance levels with 200 replicates. I used Q1 as the phenotype and converted it to binary phenotype with a prevalence 22.6%. There were 7,210 genes in each replicate, and thus 71,442,000 p-values were utilized to calculate the empirical sizes. Table 2.3 shows the empirical TIE estimates for all methods at various significance levels. Results showed that RV-TDT methods have obvious deflated TIE rates, and GEE-KM test has an inflated TIE rate. The other methods seem to preserve the nominal significance levels. Figure 2.1 shows quantile-quantile (QQ) plots, and the estimated genomic inflation factor, λ, for all methods. All results from 200 replicates were combined and were utilized to build QQ plots. Figure 2.1 shows that PedCMC, PedGene, and *FARVAT* seem to control the TIE rates well, but the estimated inflation factors of C-alpha and SKAT-O tests from *FARVAT* show

some inflation. QQ plots of results from RV-TDT show obvious deflation, and the extent of deflation is substantial for VT-BRV, VT-CMC, and WSS. Statistics in RV-TDT handle only trio data, and it may be the main reason of the deflation. The results for GEE-KM appear to be invalid. GEE-KM used the sandwich estimators for the correlation matrix between family members, and its results can be biased if the number of repeated measurement is not sufficient (Morel et al. 2003). In our case, family sizes are different, and thus the sandwich estimator was estimated with a single observation, which may be the main reason of the invalid results from GEE-KM.

Genes with the top 6 largest effects on both simulated SBP and DBP were selected to evaluate the empirical powers for all the methods. Rare variants in the selected genes with causal effects on SBP and DBP are all included for each gene-set file, and a certain number of rare variants with no effect in each gene were randomly selected to make the proportion of causal variants 10%, 25% and 50%. The empirical powers for RV-TDT are all zero, and thus are not presented in Table 2.4. Table 2.4 shows that the *FARVAT* method seems to be the most efficient and it is followed by PedGene, though the differences are small. In particular, the statistical efficiency of burden and C-alpha/kernel statistics depends on the unknown disease model, and the empirical power estimates of the SKAT-O-type *FARVAT* are usually close to the most efficient approaches. Therefore, the robust statistic against unknown genetic distributions of causal variants is uniquely provided by *FARVAT*. Power when 50% of rare variants are causal are less than those when 10% are causal, which

might be attributed to insufficient number of replicates. Overall, I can conclude that *FARVAT* and PedGene are usually the most efficient methods for the rare variant analysis with extended families, and the SKAT-O test provided by *FARVAT* is a robust method under different disease models.

Furthermore, I compared other features of each method, such as computational time, and the summary is provided in Table 2.5. According to Table 2.5, GEE-KM is a unique statistic for prospective design, and it compares the phenotypic distributions for each coded genotype while the other methods compare genetic distributions between affected and unaffected individuals. GEE-KM and PedGene can adjust effect of covariates with a logistic link function. *FARVAT* utilizes the linear mixed model to adjust the effect of covariates. Work by Crowder (Crowder 1985, Crowder 1987) suggests that the choice of a linear mixed model often work reasonably well for dichotomous phenotypes. The SKAT-O-type statistic which is robust against the distribution of genetic effects is uniquely provided by *FARVAT*. Last, in our analyses, I used Intel (R) Xeon (R) CPU E5-2620 0 @ 2.00GHz with 10 node and 80 gigabyte memory, and computational time to complete all analyses is shown. The computational time difference is related with the programming language, and software implemented with C/C++ is usually fast (Lee et al. 2012). Table 2.5 shows that *FARVAT* is the most computationally efficient.

Table 2.3 Empirical sizes calculated with 7,210 genes from 200 replicates.

α	RV-TDT					GEE-KM	PedCMC	PedGene		FARVAT		
	CMC	BRV	VT-BRV	VT-CMC	WSS			Kernel	Burden	C-alpha	Burden	SKAT-O
0.1	0.0108	0.0130	0	0	0	0.2137	0.0714	0.0895	0.0879	0.0865	0.0888	0.0864
0.05	0.0040	0.0040	0	0	0	0.1050	0.0357	0.0490	0.0433	0.0445	0.0434	0.0450
0.01	0.0009	0.0009	0	0	0	0.0163	0.0079	0.0141	0.0098	0.0112	0.0092	0.0115
0.005	0.0004	0.0004	0	0	0	0.0066	0.0043	0.0086	0.0056	0.0065	0.0050	0.0068
0.001	0	0	0	0	0	0.0006	0.0011	0.0029	0.0017	0.0020	0.0013	0.0021

The definition of the acronyms in Table 2.3: 1) α: significance level; 2) CMC: the combined multivariate and collapsing method; 3) BRV: the burden test of rare variants; 4) VT-BRV: the burden test of rare variants with variable threshold; 5) VT-CMC: the combined multivariate and collapsing method with variable threshold; 6) WSS: the weighted-sum statistic test; 7) Burden: the burden test; 8) Kernel: the kernel test, a type of variance component test; 9) C-alpha: the C-alpha score test, a type of variance component test; 10) SKAT-O: the optimal sequence kernel association test.

Table 2.4 Empirical power for the top 6 causal genes affecting both simulated SBP and DBP at the 0.05 significant level.

GENE	Proportion of causal variants	GEE-KM	PedCMC	PedGene		*FARVAT*		
				Kernel	Burden	C-alpha	Burden	SKAT-O
MAP4	10%	0.005	0.110	0.065	0.015	0.160	0.055	0.105
	50%	0.075	0.165	0.190	0.485	0.270	0.545	0.435
NRF1	10%	0.010	0.000	0.005	0.010	0.015	0.020	0.020
	50%	0.005	0.020	0.115	0.065	0.070	0.015	0.055
TNN	10%	0.045	0.005	0.005	0.005	0.005	0.010	0.005
	50%	0.085	0.020	0.025	0.020	0.025	0.025	0.025
LEPR	10%	0.010	0.075	0.005	0.045	0.010	0.055	0.030
	50%	0.000	0.010	0.020	0.010	0.020	0.020	0.010
FLT3	10%	0.000	0.245	0.440	0.160	0.505	0.255	0.450
	50%	0.035	0.040	0.525	0.410	0.450	0.395	0.425
ZNF443	10%	0.215	0.005	0.090	0.090	0.060	0.065	0.050
	50%	0.185	0	0.190	0.045	0.125	0.010	0.075
Mean	10%	0.048	0.073	0.102	0.054	0.126	0.077	0.110
	50%	0.064	0.043	0.178	0.173	0.160	0.168	0.171
Median	10%	0.010	0.040	0.035	0.030	0.038	0.055	0.040
	50%	0.055	0.020	0.153	0.055	0.098	0.023	0.065

The definition of the acronyms in Table 2.4: 1) Burden: the burden test; 2) Kernel: the kernel test, a type of variance component test; 3) C-alpha: the C-alpha score test, a type of variance component test; 4) SKAT-O: the optimal sequence kernel association test.

Table 2.5 Summary for the selected methods.

Method	Design	Phenotype		Statistic			Covariate adjustment	Language	Computing Time (hour)
		quantitative	dichotomous	Burden	VC	SKAT-O			
RV-TDT	retrospective		√	√				C	20
GEE-KM	prospective	√	√		√		√	R	40
PedCMC	retrospective		√	√				C	1.7
PedGene	retrospective		√	√	√		√	R	40
FARVAT	retrospective	√	√	√	√	√	√	C	1.7

The definition of the acronyms in Table 2.5: 1) Burden: the burden test; 2) VC: the variance component test; 3) SKAT-O: the optimal sequence kernel association test.

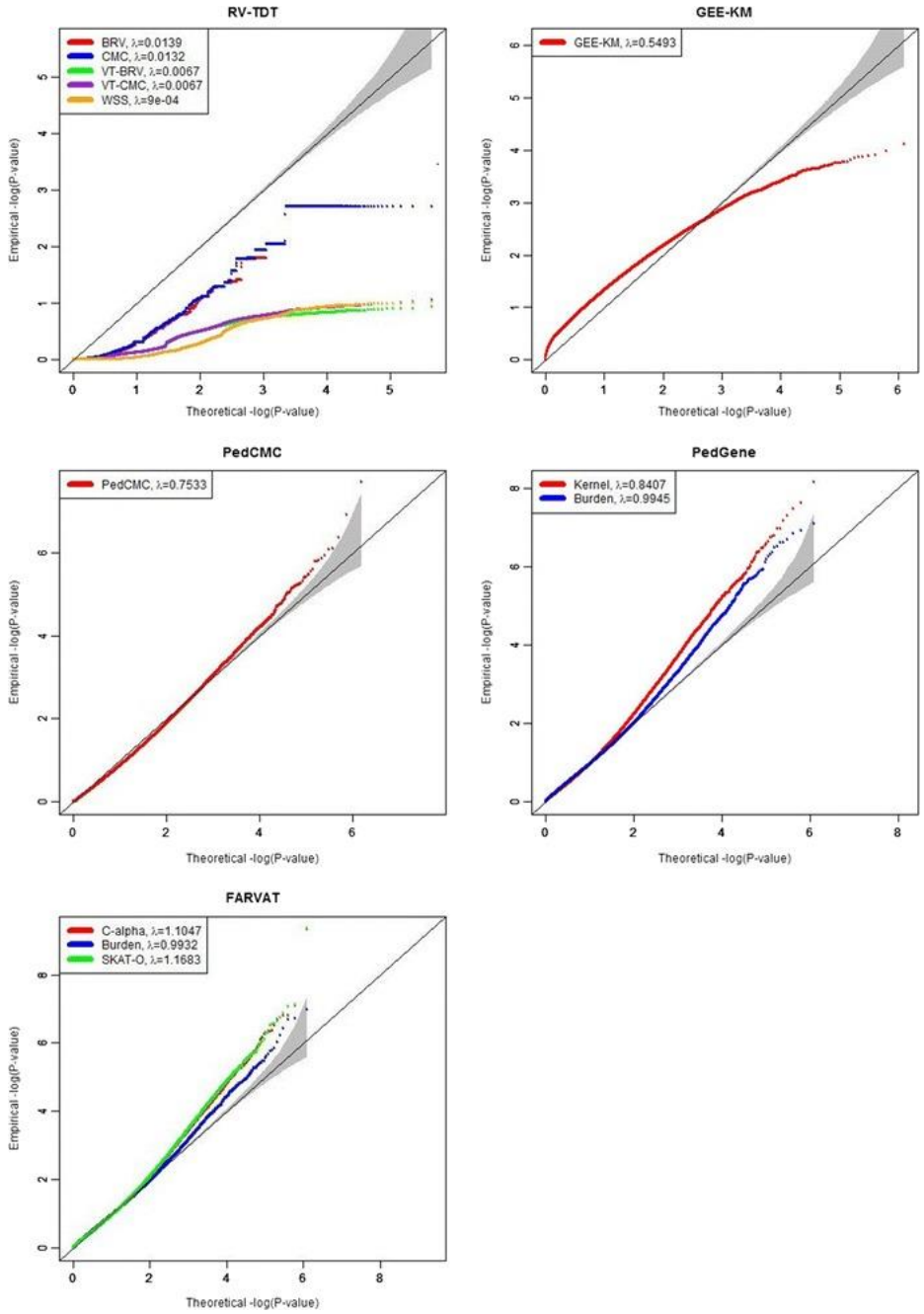


Figure 2.1 Quantile–quantile (QQ) plots for all methods

2.3 Conclusions

In this chapter, I evaluated several family-based association tests for detecting rare variants using GAW19 data. I found that *FARVAT* and PedGene usually provide similar statistical efficiency, and recommend the SKAT-O statistic provided by *FARVAT* because its power has been robust under various disease models. In addition, *FARVAT* can be applied to both quantitative and dichotomous phenotypes and was computationally fast because it was implemented with C++. Furthermore, it can load various input file formats, and provides additional information about MACs. I concluded that *FARVAT* is a good strategy for rare variant association tests with extended families in terms of both computational and statistical efficiency.

This chapter was published in *Genetic Epidemiology*
as a partial fulfillment of Longfei Wang's PhD program.

Chapter 3

Family-based Rare Variant Association Test for Multivariate Phenotypes

3.1 Introduction

mFARVAT is a quasi-likelihood-based score test for rare variant association analysis with multiple phenotypes, and tests both homogeneous and heterogeneous effects of each variant on multiple phenotypes. The method can analyze both quantitative and dichotomous phenotypes, and is robust against population substructure if the correlation matrix between individuals can be estimated from large-scale genetic data. *mFARVAT* is implemented in C++, and is computationally fast even for extended families. Simulation results show that the proposed method is generally robust and efficient for various disease

models. Furthermore, *mFARVAT* was applied to multiple phenotypes associated with COPD, and some promising results illustrate its practical value.

3.2 Methods

3.2.1 Notations and the disease model

For genetic association analyses either prospective or retrospective approaches can be selected and the choice of strategy depends on the sampling scheme. However, it has been shown that even for prospectively selected samples, retrospective analyses can preserve virtually similar statistical power as prospective analyses. Additionally, retrospective strategies are robust against non-normality of phenotypes, and are computationally less intensive (Won and Lange 2013). Therefore, I consider retrospective analysis for both prospectively and retrospectively selected samples, and genetic association is detected by testing the independence of genotype distributions with phenotypes.

Association between M genetic variants and Q phenotypes is examined, and I denote the coded genotype of individual j in family i at variant m and phenotype q by g_{ijm} and y_{ijq}, respectively. I assume there are n families and n_i individuals in family i. Thus, the sample size, N, is $\sum_{i=1}^{n} n_i$. I let

$$\mathbf{G}^m = \begin{bmatrix} g_{11m} \\ \vdots \\ g_{nn_nm} \end{bmatrix}, \mathbf{G} = (\mathbf{G}^1, \dots, \mathbf{G}^M), \text{ and}$$

$$\mathbf{Y}^q = \begin{bmatrix} y_{11q} \\ \vdots \\ y_{nn_nq} \end{bmatrix}, \mathbf{Y} = (\mathbf{Y}^1, \ldots, \mathbf{Y}^Q).$$

I also define

$$\mathbf{G}_{ij} = \begin{bmatrix} g_{ij1} \\ \vdots \\ g_{ijM} \end{bmatrix}, \text{ and } \mathbf{Y}_{ij} = \begin{bmatrix} y_{ij1} \\ \vdots \\ y_{ijQ} \end{bmatrix}.$$

The genetic variance-covariance matrix between individuals can be parameterized with the kinship coefficient matrix (KCM), $\mathbf{\Phi}$. If I let ϕ_{ijk} be the kinship coefficient between individual *j* and individual *k* in family *i*, and let h_{ij} be the inbreeding coefficient for individual *j* in family *i*,

$$\mathbf{\Phi}_i = \begin{bmatrix} 1 + h_{i1} & \cdots & 2\phi_{i1n} \\ \vdots & \ddots & \vdots \\ 2\phi_{in1} & \cdots & 1 + h_{in} \end{bmatrix},$$

and I define

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_1 & 0 & 0 & \cdots \\ 0 & \mathbf{\Phi}_2 & 0 & \cdots \\ 0 & 0 & \mathbf{\Phi}_3 & \ddots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix}.$$

In the presence of population substructure, $\mathbf{\Phi}$ should be replaced with the genetic relationship matrix (GRM) to provide statistically valid results (Thornton et al. 2012). The variance-covariance matrix between the *M* additively coded markers is denoted by $\mathbf{\Sigma}$, and I assume that

$$\text{cov}\left(\mathbf{G}_{ij}, \mathbf{G}_{i'j'}\right) \approx 2\phi_{ij,i'j'}\text{var}\left(\mathbf{G}_{ij}\right) = 2\phi_{ij,i'j'}\mathbf{\Sigma}.$$

Then I can easily show that

$$\mathrm{var}\big(\mathrm{vec}(\mathbf{G})\big) \approx \boldsymbol{\Sigma}\otimes\boldsymbol{\Phi}.$$

3.2.2 Choice of offset

It has been shown that the statistical efficiency of test statistics in retrospective analysis can be improved by adjusting phenotypes for relevant covariates (Lange et al. 2002). For our score statistic, I introduced a new parameter μ_{ijq} for phenotype *q* of individual *j* in family *i*, which will be called the offset in the remainder of this chapter (Won and Lange 2013). I set

$$\boldsymbol{\mu}_{ij} = \begin{bmatrix} \mu_{ij1} \\ \vdots \\ \mu_{ijQ} \end{bmatrix}, \boldsymbol{\mu} = \left(\boldsymbol{\mu}_{11}^{t}, \ldots, \boldsymbol{\mu}_{\boldsymbol{n}n_n}^{t}\right)^{t}, \mathbf{T}_{ij} = \mathbf{Y}_{ij} - \boldsymbol{\mu}_{ij}, \mathbf{T} = \mathbf{Y} - \boldsymbol{\mu}.$$

Statistical efficiency depends on $\boldsymbol{\mu}$, and thus its elements need to be carefully selected. The offset $\boldsymbol{\mu}$ can be either calculated by the BLUP with covariates, as done for SKAT, or the disease prevalence can be used (Won and Lange 2013). The most efficient $\boldsymbol{\mu}$ will depend on the sampling scheme. If families are randomly selected, BLUP was shown to be most efficient for both dichotomous and quantitative phenotypes (Won and Lange 2013), while prevalence was recommended to study dichotomous phenotypes if families with a large number of affected family members are selected (Thornton and McPeek 2007, Won and Lange 2013). Therefore, I chose BLUP and prevalence

as offsets for quantitative phenotypes and dichotomous phenotypes, respectively.

3.2.3 Score for quasi-likelihood

I let $\mathbf{e}_{ij}$ be an N dimensional vector in which the $\left(j + \sum_{i'=1}^{i-1} n_{i'}\right)^{\text{th}}$ element is 1 and the others are 0, and $\mathbf{1}_w$ be a column vector with w elements all equal to 1. I denote the effect of rare variant m on phenotype q as β_{mq} which is the regression coefficients of the phenotype on the causal variants. I consider the score statistic and thus β_{mq} is not needed to be estimated. However, the false positive rates can be inflated and the statistic for each β_{mq} has large false negative rates. Therefore, collapsed genotype scores were utilized to prevent these problems. Under the null hypothesis, which is $\beta_{11} = \cdots = \beta_{MQ} = 0$, the best linear unbiased estimator for $E(\mathbf{G}^m)$ (McPeek et al. 2004) is

$$\mathbf{1}_N(\mathbf{1}_N^t\mathbf{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\mathbf{\Phi}^{-1}\mathbf{G}^m,$$

and if I let $\mathbf{A} = \mathbf{\Phi}^{-1} - \mathbf{\Phi}^{-1}\mathbf{1}_N(\mathbf{1}_N^t\mathbf{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\mathbf{\Phi}^{-1}$, I can define $\mathbf{S}_{ij}^m$ for the individual j in family i by

$$\mathbf{S}_{ij}^m = \left(\mathbf{T}_{ij}\mathbf{e}_{ij}^t\right)\mathbf{\Phi}\mathbf{A}\mathbf{G}^m.$$

Based on MFQLS (Won et al. 2015), the score vector for the M variants can be defined by

$$\mathbf{S} = (\mathbf{S}^1, \dots, \mathbf{S}^M) = \mathbf{T}^t\mathbf{\Phi}\mathbf{A}\mathbf{G},$$

and because $\mathrm{var}\big(\mathrm{vec}(\mathbf{G})\big) \approx \boldsymbol{\Sigma}\otimes\boldsymbol{\Phi}$, the variance-covariance matrix for $\mathbf{S}$ is approximately equal to

$$\mathrm{var}\big(\mathrm{vec}(\mathbf{S})\big) \approx \boldsymbol{\Sigma}\otimes(\mathbf{T}^t\boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}\mathbf{T}).$$

3.2.4 Homogeneous *mFARVAT*

The effects of each causal variant on a phenotype, estimated as the regression coefficients of the phenotype on the causal variants, can be in the same or different directions, and I propose two different statistics for these two scenarios. The first statistic, homogeneous *mFARVAT*, assumes that effects of each causal variant on the multiple phenotypes are in the same direction, for example, when the phenotypes are highly correlated or longitudinal. For rare variant association analysis, burden tests regress phenotypes on the sum of genotype scores over rare variants. Therefore, association of the Q phenotypes with variant m can be built by testing whether $\beta_{m1} + \ldots + \beta_{mQ} = 0$, and I can provide a statistic based on $\mathbf{1}_Q^t\mathbf{S}$.

The importance of each variant is often different and statistical efficiency can be improved by weighting each variant based on its relative importance (Madsen and Browning 2009). Relative importance is usually expressed by a function of MAF. I assume that the weight for variant m is w_m and $\mathbf{W}$ is an $M\times M$ diagonal matrix with diagonal elements w_m; I choose $w_m = \mathrm{beta}(p_m, a_1, a_2)$ proposed by Wu et al(Wu et al. 2011), where p_m is the MAF of variant m and a_1

and a_2 were set to be 1 and 25 respectively. beta(p_m, a_1, a_2) is flexible because it can accommodate a broad range of scenarios by considering different a_1 and a_2, and Wu et al found that the choices of a_1 and a_2 were often efficient. Then the scores for the burden and SKAT tests are, respectively,

$$\frac{1}{\mathbf{1}_Q^t \mathbf{T}^t \mathbf{\Phi A \Phi T} \mathbf{1}_Q} \mathbf{1}_Q^t \mathbf{S W 1}_M \mathbf{1}_M^t \mathbf{W S}^t \mathbf{1}_Q,$$

and

$$\frac{1}{\mathbf{1}_Q^t \mathbf{T}^t \mathbf{\Phi A \Phi T} \mathbf{1}_Q} \mathbf{1}_Q^t \mathbf{S W W S}^t \mathbf{1}_Q.$$

If I let

$$\mathbf{R}_\rho^{Hom} = (1-\rho)\mathbf{I}_M + \rho \mathbf{1}_M \mathbf{1}_M^t,$$

scores for burden and SKAT tests can be generalized as

$$MS_\rho^{Hom} = \frac{1}{\mathbf{1}_Q^t \mathbf{T}^t \mathbf{\Phi A \Phi T} \mathbf{1}_Q} \mathbf{1}_Q^t \mathbf{S W R}_\rho^{Hom} \mathbf{W S}^t \mathbf{1}_Q,$$

where the optimal choice of ρ depends on the distribution of rare variant effects on the multiple phenotypes.

I denote the eigenvalues of $\mathbf{\Sigma}^{1/2} \mathbf{W R}_\rho^{Hom} \mathbf{W \Sigma}^{1/2}$ by $\left(\lambda_1^\rho, \ldots, \lambda_M^\rho\right)$. If I let $\chi_{1,m}^2$ be an independent chi-square distribution with a single degree of freedom, I have

$$MS_{\rho}^{Hom} \sim \sum_{m=1}^{M} \lambda_m^{\rho} \chi_{1,m}^2.$$

If I denote the p-value for MS_{ρ}^{Hom} by pMS_{ρ}^{Hom}, and let $pmFARVAT_S^{Hom} = pMS_0^{Hom}$ and $pmFARVAT_B^{Hom} = pMS_1^{Hom}$, the SKAT-O *mFARVAT* (*mFARVAT*$_O$) statistic is defined by

$$mFARVAT_O^{Hom} = \min\{pMS_0^{Hom}, pMS_{0.1^2}^{Hom}, \ldots, pMS_{0.5^2}^{Hom}, pMS_1^{Hom}\}.$$

Its p-value will be denoted as $pmFARVAT_O^{Hom}$ in the remainder of this chapter, and can be calculated from the numerical algorithm for SKAT-O (Lee et al. 2012), with a small modification. If I let

$$\mathbf{Z} = \mathbf{\Sigma}^{1/2}\mathbf{W}, \text{and } \bar{\mathbf{Z}} = \mathbf{Z}\mathbf{1}_M(\mathbf{1}_M^t\mathbf{1}_M)^{-1},$$

the projection matrix onto a space spanned by $\bar{\mathbf{Z}}$ becomes $\mathbf{\Pi} = \bar{\mathbf{Z}}(\bar{\mathbf{Z}}^t\bar{\mathbf{Z}})\bar{\mathbf{Z}}^t$.

If I let

$$\mathbf{u} = \mathbf{\Sigma}^{-1/2}\mathbf{S}^t\mathbf{1}_Q \frac{1}{\sqrt{\mathbf{1}_Q^t\mathbf{T}^t\mathbf{\Phi A \Phi T}\mathbf{1}_Q}}, \mathbf{u} \sim MVN(0, \mathbf{I}_M),$$

MS_{ρ}^{Hom} becomes

$$MS_{\rho}^{Hom} = \mathbf{u}^t\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{u} = (1-\rho)\mathbf{u}^t\mathbf{Z}\mathbf{Z}^t\mathbf{u} + \rho M^2\mathbf{u}^t\bar{\mathbf{Z}}\bar{\mathbf{Z}}^t\mathbf{u}.$$

As was shown by Lee et al (Lee et al. 2012), if I let

$$\tau(\rho) = M^2\rho\bar{\mathbf{Z}}^t\bar{\mathbf{Z}} + \frac{(1-\rho)}{\bar{\mathbf{Z}}^t\bar{\mathbf{Z}}}\bar{\mathbf{Z}}^t\mathbf{Z}\mathbf{Z}^t\bar{\mathbf{Z}},$$

I have

$$MS_{\rho}^{Hom} = (1-\rho)\mathbf{u}^t(\mathbf{I}_M - \mathbf{\Pi})\mathbf{Z}\mathbf{Z}^t(\mathbf{I}_M - \mathbf{\Pi})\mathbf{u} + 2(1-\rho)\mathbf{u}^t(\mathbf{I}_M - \mathbf{\Pi})\mathbf{Z}\mathbf{Z}^t\mathbf{\Pi}\mathbf{u} + \tau(\rho)\mathbf{u}^t\mathbf{\Pi}\mathbf{u},$$

where $\mathbf{u}^t(\mathbf{I}_M - \mathbf{\Pi})\mathbf{Z}\mathbf{Z}^t(\mathbf{I}_M - \mathbf{\Pi})\mathbf{u}$, $\mathbf{u}^t(\mathbf{I}_M - \mathbf{\Pi})\mathbf{Z}\mathbf{Z}^t\mathbf{\Pi}\mathbf{u}$ and $\mathbf{u}^t\mathbf{\Pi}\mathbf{u}$ are mutually independent. Therefore, if I let

$$P_{min} = min\{pMS_0^{Hom}, pMS_{0.1^2}^{Hom}, \dots, pMS_{0.5^2}^{Hom}, pMS_1^{Hom}\},$$

then I have

$$P\left(MS_{\rho_0}^{Hom} \le Q_{\rho_0}(P_{\min}), \dots, MS_{\rho_L}^{Hom} \le Q_{\rho_L}(P_{\min})\right)$$

$$= E\{P\left(MS_{\rho_0}^{Hom} \le Q_{\rho_0}(P_{\min}), \dots, \mathrm{MS}_{\rho_L}^{Hom} \le Q_{\rho_L}(P_{\min}) \,|\mathbf{u}^t\mathbf{\Pi}\mathbf{u} = \eta\right)\}.$$

Conditional probability can be numerically calculated as was suggested by Lee et al (Lee et al. 2012):

$$P\left(MS_{\rho_0}^{Hom} \le Q_{\rho_0}(P_{\min}), \dots, MS_{\rho_L}^{Hom} \le Q_{\rho_L}(P_{\min}) \,|\mathbf{u}^t\mathbf{\Pi}\mathbf{u} = \eta\right).$$

3.2.5 Heterogeneous *mFARVAT*

The effects of each variant on multiple phenotypes can be heterogeneous in certain situations, and it may be reasonable to consider such effects separately.

I assumed the effects β_{mq} of variant m on multiple phenotypes follow an arbitrary distribution with mean 0 and variance τ_m. Therefore, I can provide statistics based on vec($\mathbf{S}$), and, under the null hypothesis $\beta_{11} = \cdots = \beta_{MQ} = 0$, I have

$$\mathrm{E}\{\mathrm{vec}(\mathbf{S})\} = \mathbf{0} \quad \text{and} \quad \mathrm{var}\{\mathrm{vec}(\mathbf{S})\} = \mathbf{\Sigma} \otimes \mathbf{T}^t \mathbf{\Phi} \mathbf{A} \mathbf{\Phi} \mathbf{T}.$$

If I assume that $\mathbf{I}_w$ is a $w \times w$ identity matrix and

$$\mathbf{R}_\rho^{Het} = (1-\rho)\mathbf{I}_{MQ} + \rho \mathbf{1}_{MQ} \mathbf{1}_{MQ}^t,$$

I define the generalized score by

$$MS_\rho^{Het} = \mathrm{vec}(\mathbf{S})^t \big(\mathbf{I}_Q \otimes \mathbf{W}\big) \mathbf{R}_\rho^{Het} \big(\mathbf{I}_Q \otimes \mathbf{W}\big) \mathrm{vec}(\mathbf{S}).$$

Then the burden and SKAT tests can be expressed as

$$MS_1^{Het} = \mathrm{vec}(\mathbf{S})^t \big(\mathbf{I}_Q \otimes \mathbf{W}\big) \mathbf{1}_{MQ} \mathbf{1}_{MQ}^t \big(\mathbf{I}_Q \otimes \mathbf{W}\big) \mathrm{vec}(\mathbf{S}),$$

$$MS_0^{Het} = \mathrm{vec}(\mathbf{S})^t \big(\mathbf{I}_Q \otimes \mathbf{W}\big) \big(\mathbf{I}_Q \otimes \mathbf{W}\big) \mathrm{vec}(\mathbf{S}).$$

If I let $\left(\lambda_1'^{\rho}, \ldots, \lambda_{MQ}'^{\rho}\right)$ be the eigenvalues of

$$\big(\mathbf{\Sigma}^{1/2} \otimes (\mathbf{T}^t \mathbf{\Phi} \mathbf{A} \mathbf{\Phi} \mathbf{T})^{1/2}\big) \big(\mathbf{I}_Q \otimes \mathbf{W}\big) \mathbf{R}_\rho$$
$$\times \big(\mathbf{I}_Q \otimes \mathbf{W}\big) \big(\mathbf{\Sigma}^{1/2} \otimes (\mathbf{T}^t \mathbf{\Phi} \mathbf{A} \mathbf{\Phi} \mathbf{T})^{1/2}\big),$$

then I have

$$MS_{\rho}^{Het} \sim \sum_{l=1}^{MQ} \lambda_l'^{\rho} \chi_{1,l}^2 \text{ under } H_0.$$

P-values for MS_{ρ}^{Het} will be denoted by pMS_{ρ}^{Het}, and I let $pmFARVAT_S^{Het} = pMS_0^{Het}$ and $pmFARVAT_B^{Het} = pMS_1^{Het}$. I consider

$$mFARVAT_O^{Het} = \min\{pMS_0^{Het}, pMS_{0.1^2}^{Het}, \dots, pMS_{0.5^2}^{Het}, pMS_1^{Het}\}.$$

I let the p-value for $mFARVAT_O^{Het}$ be $pmFARVAT_O^{Het}$ and the detailed algorithm to calculate the asymptotic p-value is provided in the next section. Similarly, for $mFARVAT_O^{Het}$, I assume

$$\mathbf{Z} = \text{var}\big(\text{vec}(\mathbf{S})\big)^{1/2}\big(\mathbf{I}_Q \otimes \mathbf{W}\big), \text{and } \bar{\mathbf{Z}} = \mathbf{Z}\mathbf{1}_{MQ}\big(\mathbf{1}_{MQ}^t\mathbf{1}_{MQ}\big)^{-1}.$$

Then the projection matrix on a space spanned by $\bar{\mathbf{Z}}$ is $\mathbf{\Pi} = \bar{\mathbf{Z}}(\bar{\mathbf{Z}}^t\bar{\mathbf{Z}})\bar{\mathbf{Z}}^t$.

If I let

$$\mathbf{u} = \text{var}\big(\text{vec}(\mathbf{S})\big)^{-1/2}\text{vec}(\mathbf{S}), \mathbf{u} \sim MVN\big(0, \mathbf{I}_{MQ}\big),$$

MS_{ρ}^{Het} becomes

$$MS_{\rho}^{Het} = \mathbf{u}^t\text{var}\big(\text{vec}(\mathbf{S})\big)^{\frac{1}{2}}\big(\mathbf{I}_Q \otimes \mathbf{W}\big)\text{var}\big(\text{vec}(\mathbf{S})\big)^{\frac{1}{2}}\mathbf{u}$$

$$= (1-\rho)\mathbf{u}^t\mathbf{Z}\mathbf{Z}^t\mathbf{u} + \rho(MQ)^2\mathbf{u}^t\bar{\mathbf{Z}}\bar{\mathbf{Z}}^t\mathbf{u}.$$

As was suggested by Lee et al(Lee et al. 2012), if I let

$$\tau(\rho) = (MQ)^2\rho\bar{\mathbf{Z}}^t\bar{\mathbf{Z}} + \frac{(1-\rho)}{\bar{\mathbf{Z}}^t\bar{\mathbf{Z}}}\bar{\mathbf{Z}}^t\mathbf{Z}\mathbf{Z}^t\bar{\mathbf{Z}},$$

I have

$$MS_{\rho}^{Het} = (1-\rho)\mathbf{u}^t\left(\mathbf{I}_{MQ} - \mathbf{\Pi}\right)\mathbf{Z}\mathbf{Z}^t\left(\mathbf{I}_{MQ} - \mathbf{\Pi}\right)\mathbf{u}$$

$$+ 2(1-\rho)\mathbf{u}^t\left(\mathbf{I}_{MQ} - \mathbf{\Pi}\right)\mathbf{Z}\mathbf{Z}^t\mathbf{\Pi}\mathbf{u} + \tau(\rho)\mathbf{u}^t\mathbf{\Pi}\mathbf{u},$$

Therefore, if I let $P_{\min} = min\{pMS_0^{Het}, pMS_{0.1^2}^{Het}, \dots, pMS_{0.5^2}^{Het}, pMS_1^{Het}\}$, I have

$$P\left(MS_{\rho_0}^{Het} \le Q_{\rho_0}(P_{\min}), \dots, MS_{\rho_L}^{Het} \le Q_{\rho_L}(P_{\min})\right)$$

$$= E\{P\left(MS_{\rho_0}^{Het} \le Q_{\rho_0}(P_{\min}), \dots, MS_{\rho_L}^{Het} \le Q_{\rho_L}(P_{\min}) \,|\mathbf{u}^t\mathbf{\Pi}\mathbf{u} = \eta\right)\}.$$

$P\left(MS_{\rho_0}^{Het} \le Q_{\rho_0}(P_{\min}), \dots, MS_{\rho_L}^{Het} \le Q_{\rho_L}(P_{\min}) \,|\mathbf{u}^t\mathbf{\Pi}\mathbf{u} = \eta\right)$ can be calculated as in (Lee et al. 2012).

3.3 Simulation study

3.3.1 The simulation model

To evaluate *mFARVAT*, I simulated large families that extend three generations and consist of 10 members (see Figure 3.1). 5,000 haplotypes with 50,000 base pairs were generated under a coalescent model using the software COSI (Schaffner et al. 2005). Each haplotype was generated by setting the mutation rate at 1.5×10^{-8}. Haplotypes were randomly chosen with replacement to build founder genotypes. Nonfounder haplotypes were determined in Mendelian fashion from pairs of parents under the assumption of no recombination. For each simulated haplotype, I defined variants with sample MAFs less than 0.01 as being rare, and 60 rare variants were randomly selected.

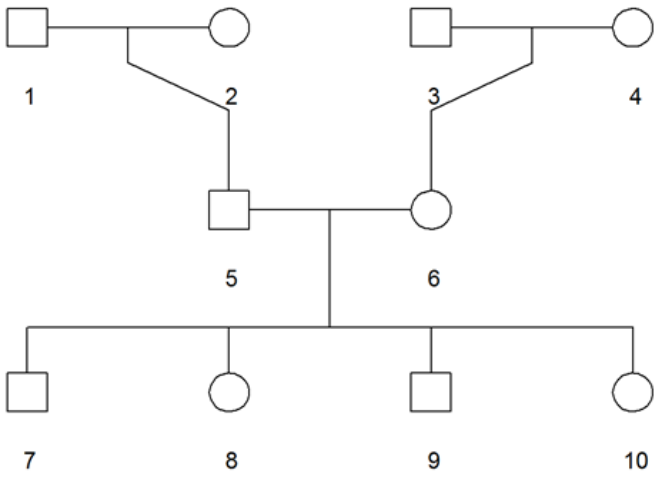


Figure 3.1 Extended family used in the simulation studies

Phenotypes were generated under the null and alternative hypotheses, and I considered both quantitative and dichotomous phenotypes. Quantitative phenotypes were defined by summing the phenotypic mean, polygenic effect, main genetic effect and random error, and I assumed there was no environmental effect shared between family members. Phenotypic means were denoted by $\alpha_{1,\ldots}$, $\alpha_{Q\text{-}1}$ and α_Q. I assumed that $\alpha_1 = 0$, $\alpha_2 = 0.3$ for $Q = 2$, and $\alpha_1 = \alpha_2 = \alpha_3 = 0$, $\alpha_4 = \alpha_5 = 0.3$ for $Q = 5$. The polygenic effects for the Q phenotypes for each founder were independently generated from $MVN(\mathbf{0},\mathbf{V_B})$, and for nonfounders the average of maternal and paternal polygenic effects were combined with values independently sampled from $MVN(\mathbf{0}, 0.5\mathbf{V_B})$. Random errors for the Q phenotypes were assumed to be independent, so the random error for phenotype q was independently sampled from $N(\mathbf{0}, \sigma_{\mathbf{E},q}^2)$. If $Q = 2$, I assumed that

$$\mathbf{V_B} = \begin{bmatrix} 1 & \sqrt{2}c \\ \sqrt{2}c & 2 \end{bmatrix}, \sigma_{E,1}^2 = 1, \sigma_{E,2}^2 = 2,$$

and if $Q = 5$, they were

$$\mathbf{V_B} = \begin{bmatrix} 1 & c & \sqrt{2}c & \sqrt{2}c & \sqrt{2}c \\ c & 1 & \sqrt{2}c & \sqrt{2}c & \sqrt{2}c \\ \sqrt{2}c & \sqrt{2}c & 2 & 2c & 2c \\ \sqrt{2}c & \sqrt{2}c & 2c & 2 & 2c \\ \sqrt{2}c & \sqrt{2}c & 2c & 2c & 2 \end{bmatrix},$$

$$\sigma_{E,1}^2 = 1, \sigma_{E,2}^2 = 2, \sigma_{E,3}^2 = 3, \sigma_{E,4}^2 = 4, \sigma_{E,5}^2 = 5.$$

For c I chose 0.5 and 0.8.

The genetic effect at variant m for phenotype q was the product of β_{mq} and the number of disease susceptibility alleles. Under the null hypothesis, β_{mq} was assumed to be 0. Under the alternative hypothesis, if I let h_a^2 be the proportion of variance explained by rare variants, β_{mq} was sampled from $U(0,v_q)$, where

$$v_q = \sqrt{\frac{\left(\sigma_{B,q}^2 + \sigma_{E,q}^2\right)h_a^2}{(1-h_a^2)\sum_{m=1}^{M}\beta_{mq}^2 2p_m(1-p_m)}}.$$

Here $\sigma_{\mathbf{B},q}^2$ indicates the $(q,q)^{\text{th}}$ element of $\mathbf{V_B}$, and I assumed that $h_a^2 = 0.02$. β_{mq} was generated for both heterogeneous and homogeneous scenarios. For homogeneous scenarios, I assumed that the effects of each rare variant on different phenotypes are similar. For example, the ratios between $\beta_{m1}, \ldots,$ and β_{mQ} were assumed to be 1:0.9 if $Q = 2$, and 1:0.9:0.8:0.7:0.6 if $Q = 5$. For heterogeneous scenario, the effects of each rare variant on phenotypes were independently generated from $U(0,v_q)$.

Simulation of dichotomous phenotypes was performed using the liability threshold model. Once the quantitative phenotypes with genetic effect, polygenic effect and random error were generated, they were transformed to being affected for quantitative phenotypes larger than the threshold, and otherwise were transformed to unaffected. The threshold was chosen to preserve the assumed disease prevalence. I assumed that prevalence of the multiple phenotypes were 0.1 or 0.2 if $Q = 2$, and 0.1, 0.2, 0.2, 0.3, or 0.3 if $Q = 5$. To allow for the ascertainment bias of dichotomous phenotypes in our

simulation studies, I assumed that families with at least one affected individual were selected for analysis.

3.3.2 Evaluation of *mFARVAT* with simulated data

To evaluate statistical validity, TIE estimates for both dichotomous and quantitative phenotypes were calculated at various significance levels using 20,000 replicates of two hundred extended families, so that each replicate sample contained 2,000 individuals. Table 3.1 shows empirical TIE estimates for homogeneous *mFARVAT* (*mFARVAT*^{Hom}) and heterogeneous *mFARVAT* (*mFARVAT*^{Het}) at the 0.05, 0.01, 0.001, and 2.5×10^{-6} significance levels. The estimates are virtually equal to the nominal significance levels for both quantitative and dichotomous phenotypes. Quantile-quantile (QQ) plots in Figures 3.2 -3.5 also show consistent results, and I conclude that *mFARVAT*^{Het} and *mFARVAT*^{Hom} are statistically valid.

Table 3.1 Type I error estimates from the simulation study. The empirical type I error was estimated for heterogeneous and homogeneous SKAT-O type *mFARVAT* with 20,000 replicates at the 0.05, 0.01, 0.001 and 2.5×10^{-6} significance levels. I assumed that the number of rare variants is 60, and that their MAFs were generated as $U(0,v_q)$.

Correlation	Type	*Q*	*mFARVAT*^{Het}				*mFARVAT*^{Hom}			
			0.05	**0.01**	**0.001**	**2.5E-6**	**0.05**	**0.01**	**0.001**	**2.5E-6**
0.5	Quantitative	2	0.0449	0.0087	0.0009	0	0.0470	0.0082	0.0009	0
		5	0.0481	0.0098	0.0009	0	0.0502	0.0084	0.0009	0
	Dichotomous	2	0.0503	0.0110	0.0010	5e-5	0.0502	0.0106	0.0008	0
		5	0.0502	0.0083	0.0013	0	0.0483	0.0093	0.0012	0
0.8	Quantitative	2	0.0443	0.0087	0.0007	0	0.0466	0.0091	0.0011	0
		5	0.0491	0.0099	0.0014	0	0.0498	0.0099	0.0014	0
	Dichotomous	2	0.0505	0.0111	0.0014	0	0.0507	0.0106	0.0012	0
		5	0.0484	0.0095	0.0011	0	0.0487	0.0089	0.0011	0

The definition of the acronyms in Table 3.1: 1) Correlation: the correlation among the phenotypes; 2) Type: the type of phenotypes; 3) *Q*: the number of phenotypes.

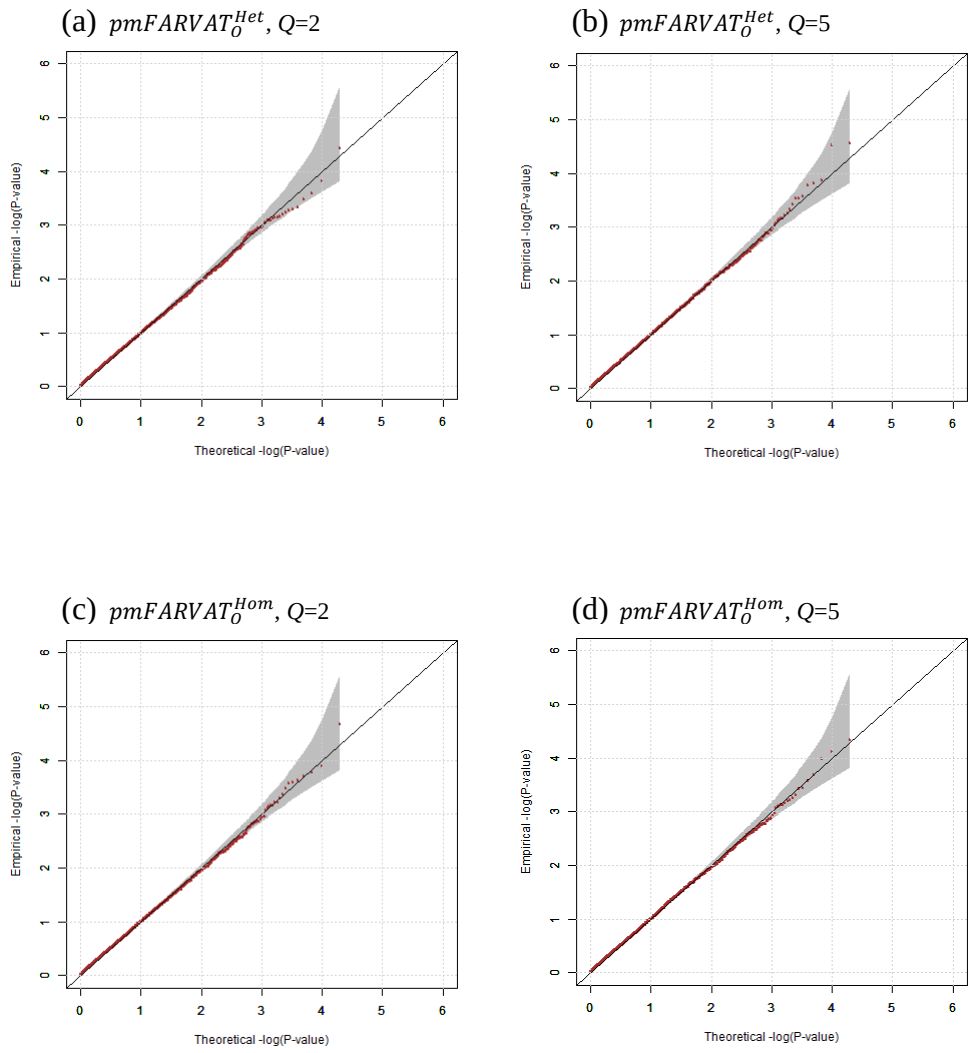


Figure 3.2 QQ plots of *mFARVAT*$_O$ for quantitative phenotypes with *c* = 0.5. The empirical p-values for heterogeneous and homogeneous *mFARVAT* were calculated under the null hypothesis with 20,000 replicates for Q=2 and Q=5. The correlation among phenotypes is 0.5.

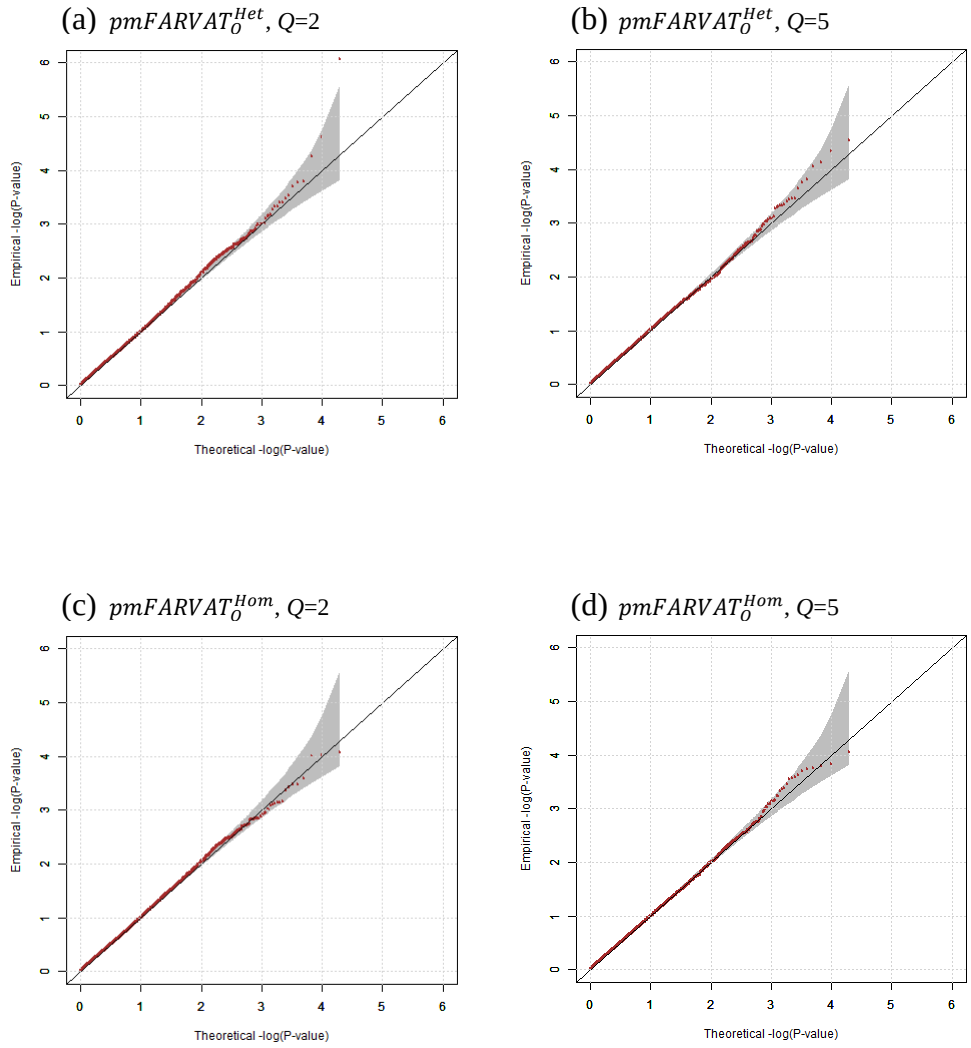


Figure 3.3 QQ plots of *mFARVAT*$_O$ for dichotomous phenotypes with c = 0.5. The empirical p-values for heterogeneous and homogeneous *mFARVAT* were calculated under the null hypothesis with 20,000 replicates for Q=2 and Q=5. The correlation among phenotypes is 0.5.

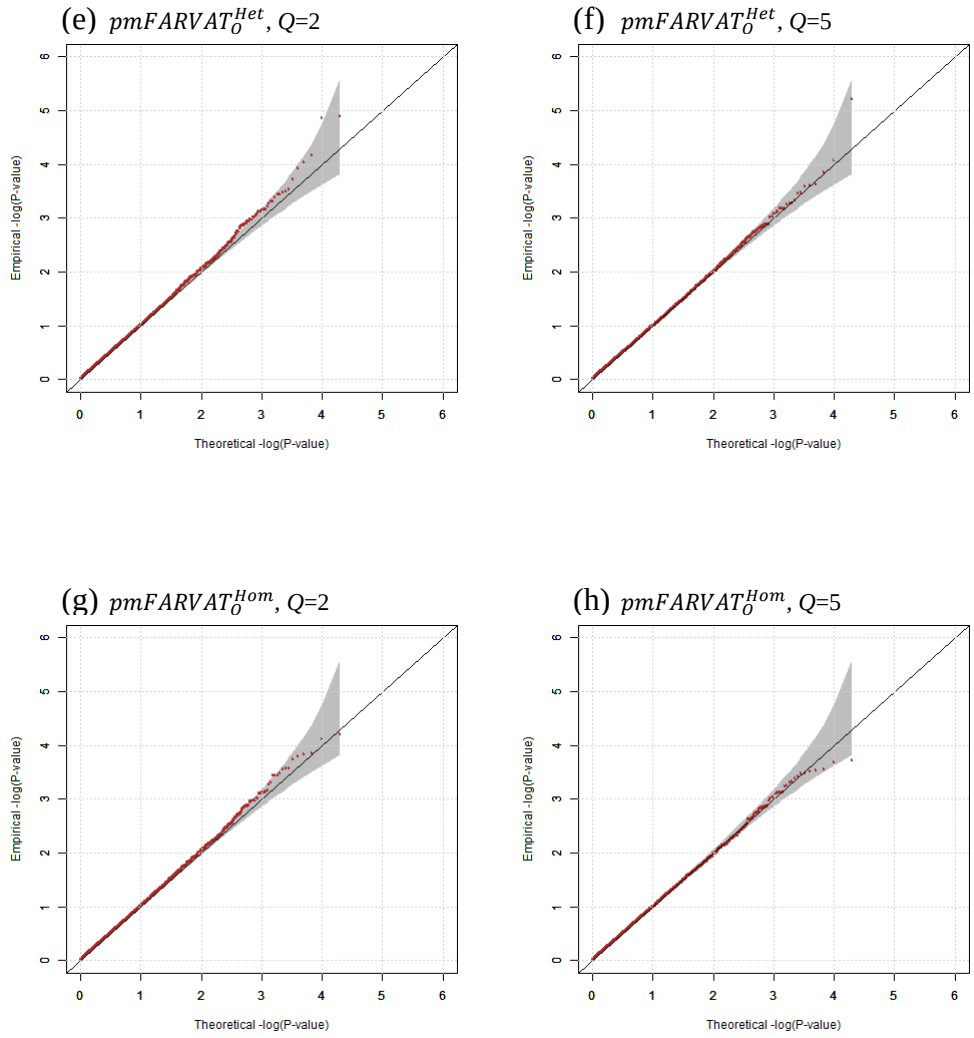


Figure 3.4 QQ plots of *mFARVAT*$_O$ for quantitative phenotypes with *c* = 0.8. The empirical p-values for heterogeneous and homogeneous *mFARVAT* were calculated under the null hypothesis with 20,000 replicates for Q=2 and Q=5. The correlation among phenotypes is 0.8.

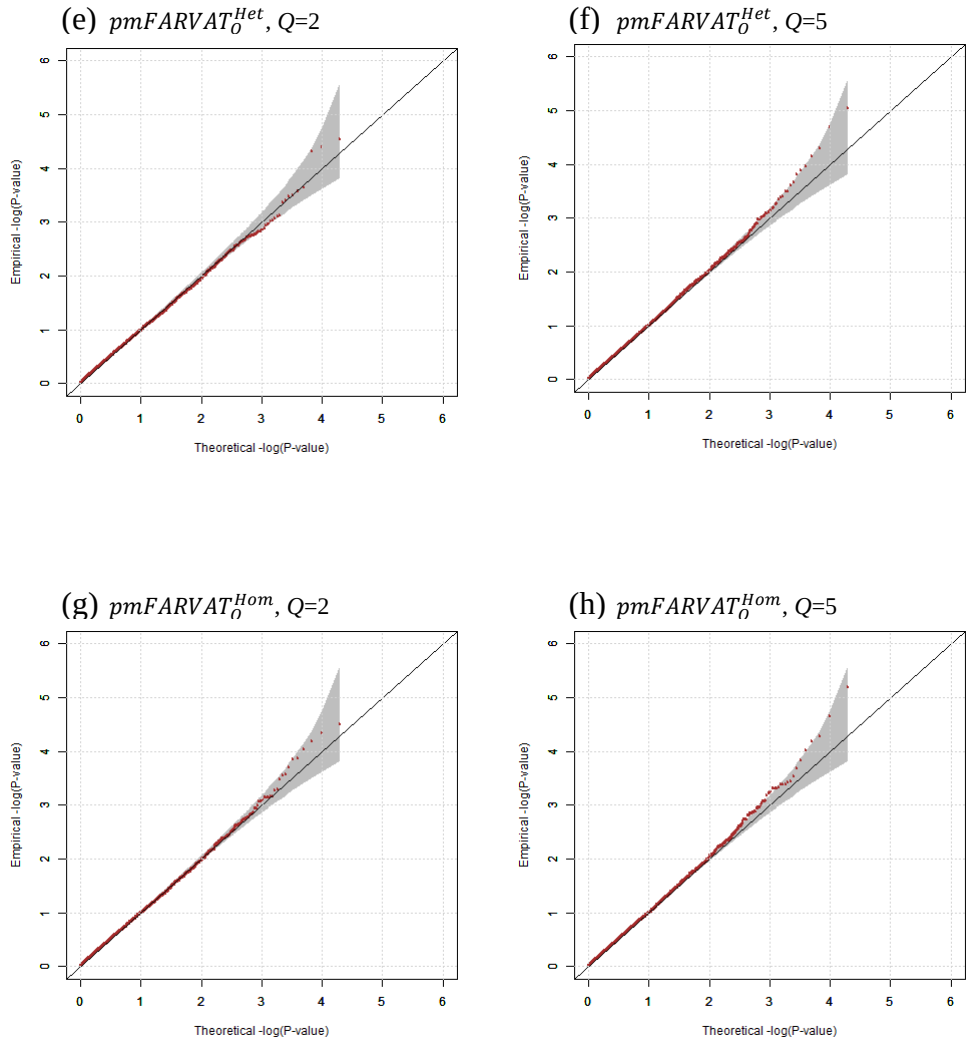


Figure 3.5 QQ plots of *mFARVAT*$_O$ for dichotomous phenotypes with *c* = 0.8. The empirical p-values for heterogeneous and homogeneous *mFARVAT* were calculated under the null hypothesis with 20,000 replicates for Q=2 and Q=5. The correlation among phenotypes is 0.8.

Empirical power estimates were calculated at the 10^{-4} significance level with correlations 0.5 and 0.8 for quantitative phenotypes (for the underlying quantitative phenotypes in the case of dichotomous phenotypes). I considered two different scenarios, in which either all or half the rare variants were causal, and assumed that 50%, 80% and 100% of causal variants were deleterious, with the rest being protective. Empirical power estimates were calculated with 2,000 replicates for six different statistics: (1) $mFARVAT_O^{Het}$; (2) $mFARVAT_O^{Hom}$; (3) $mFARVAT_S^{Het}$; (4) $mFARVAT_S^{Hom}$; (5) $mFARVAT_B^{Het}$; (6) $mFARVAT_B^{Hom}$. Results are provided in Tables 3.2-3.4 and Tables 3.5-3.7, which represent respectively scenarios where all or half the rare variants are causal. Notably, each method performed similarly in both scenarios, although the empirical power estimates improve if causal variants are more abundant.

I first examined the efficiency of the methods. Tables 3.2-3.7 confirm that the most efficient method depends on the disease model, which tends to be unknown. For example, when all the rare causal variants have deleterious effects on all phenotypes, burden *mFARVAT* (*mFARVAT*$_B$) outperforms all other approaches, but if there are variants with deleterious and protective effects, SKAT *mFARVAT* (*mFARVAT*$_S$) is the most efficient. SKAT-O *mFARVAT* (*mFARVAT*$_O$) is not always the best, but its empirical power estimates are usually very close to those of the most efficient approach. Therefore, our results are consistent with previous findings that *mFARVAT*$_O$ is robust and efficient for various disease models (Lee et al. 2012).

Table 3.2 Empirical power estimates when all rare variants are causal and 100% of them are deleterious. Empirical power of $\boldsymbol{mFARVAT_S^{Het}}$, $\boldsymbol{mFARVAT_B^{Het}}$, $\boldsymbol{mFARVAT_O^{Het}}$, $\boldsymbol{mFARVAT_B^{Hom}}$, $\boldsymbol{mFARVAT_S^{Hom}}$ and $\boldsymbol{mFARVAT_O^{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes (Q = 2 and Q = 5) with homogeneous and heterogeneous effects and different correlations (c = 0.5 and c = 0.8) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

Q	Type	c	Eff	*FARVAT*			$mFARVAT^{Het}$			$mFARVAT^{Hom}$		
				SKAT	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.208	0.712	0.738	0.331	0.908	0.912	0.337	0.896	0.900
			Hom	0.196	0.766	0.778	0.353	0.928	0.927	0.439	0.915	0.925
		0.8	Het	0.200	0.713	0.723	0.310	0.876	0.875	0.290	0.865	0.859
			Hom	0.201	0.705	0.729	0.333	0.865	0.874	0.373	0.853	0.874
	Q	0.5	Het	0.350	0.987	0.987	0.531	0.998	0.998	0.593	0.999	0.998
			Hom	0.396	0.984	0.979	0.574	0.998	0.998	0.755	0.996	0.997
		0.8	Het	0.251	0.980	0.979	0.490	0.995	0.999	0.486	0.995	0.995
			Hom	0.365	0.977	0.977	0.509	0.996	0.995	0.607	0.996	0.995

5	D	0.5	Het	0.317	0.924	0.934	0.839	1.000	1.000	0.826	1.000	1.000
			Hom	0.315	0.948	0.955	0.868	1.000	1.000	0.947	1.000	1.000
		0.8	Het	0.267	0.887	0.900	0.706	0.991	0.995	0.635	0.990	0.992
			Hom	0.265	0.893	0.914	0.756	0.995	0.995	0.814	0.995	0.995
	Q	0.5	Het	0.540	0.998	0.998	0.952	1.000	1.000	0.973	1.000	1.000
			Hom	0.602	1.000	1.000	0.968	1.000	1.000	0.999	1.000	1.000
		0.8	Het	0.495	0.992	0.993	0.879	1.000	1.000	0.836	1.000	1.000
			Hom	0.525	0.994	0.994	0.890	1.000	1.000	0.957	1.000	1.000

The definition of the acronyms in Table 3.2: 1) *Q*: the number of phenotypes; 2) Type: the type of phenotypes, D – dichotomous, Q - quantitative; 4) c: the correlation among the phenotypes; 5) Eff: the underlying genetic effects architecture, Hom – homogeneous effects, Het – heterogeneous effects; 6) SKAT: the sequence kernel association test; 7) Burden: the burden test; 8) SKAT-O: the optimal SKAT.

Table 3.3 Empirical power estimates when all rare variants are causal and 80% of them are deleterious. Empirical power of $\boldsymbol{mFARVAT_S^{Het}}$, $\boldsymbol{mFARVAT_B^{Het}}$, $\boldsymbol{mFARVAT_O^{Het}}$, $\boldsymbol{mFARVAT_B^{Hom}}$, $\boldsymbol{mFARVAT_S^{Hom}}$ *and* $\boldsymbol{mFARVAT_O^{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes (Q = 2 and Q = 5) with homogeneous and heterogeneous effects and different correlations (c = 0.5 and c = 0.8) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

Q	Type	c	Eff	*FARVAT*			$mFARVAT^{Het}$			$mFARVAT^{Hom}$		
				SKAT	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.129	0.148	0.289	0.231	0.327	0.464	0.135	0.302	0.372
			Hom	0.150	0.194	0.326	0.252	0.389	0.525	0.342	0.356	0.550
		0.8	Het	0.111	0.146	0.270	0.191	0.263	0.422	0.092	0.242	0.316
			Hom	0.133	0.164	0.283	0.212	0.313	0.447	0.264	0.279	0.462
	Q	0.5	Het	0.355	0.414	0.627	0.523	0.592	0.808	0.301	0.581	0.692
			Hom	0.368	0.467	0.670	0.546	0.678	0.854	0.718	0.660	0.892
		0.8	Het	0.331	0.376	0.608	0.451	0.491	0.736	0.190	0.492	0.561
			Hom	0.319	0.407	0.608	0.461	0.564	0.753	0.577	0.536	0.790

5	D	0.5	Het	0.214	0.270	0.479	0.707	0.763	0.903	0.272	0.745	0.764
			Hom	0.228	0.328	0.512	0.750	0.844	0.931	0.887	0.814	0.952
		0.8	Het	0.179	0.215	0.386	0.629	0.590	0.819	0.143	0.562	0.586
			Hom	0.195	0.290	0.453	0.622	0.705	0.855	0.742	0.672	0.881
	Q	0.5	Het	0.577	0.574	0.831	0.962	0.922	0.997	0.459	0.923	0.931
			Hom	0.546	0.643	0.839	0.934	0.961	0.997	0.992	0.954	1.000
		0.8	Het	0.527	0.490	0.765	0.915	0.802	0.972	0.238	0.791	0.804
			Hom	0.477	0.566	0.773	0.865	0.846	0.971	0.953	0.826	0.989

The definition of the acronyms in Table 3.3: 1) *Q*: the number of phenotypes; 2) Type: the type of phenotypes, D – dichotomous, Q - quantitative; 4) c: the correlation among the phenotypes; 5) Eff: the underlying genetic effects architecture, Hom – homogeneous effects, Het – heterogeneous effects; 6) SKAT: the sequence kernel association test; 7) Burden: the burden test; 8) SKAT-O: the optimal SKAT.

Table 3.4 Empirical power estimates when all rare variants are causal and 50% of them are deleterious. Empirical power of $\boldsymbol{mFARVAT_S^{Het}}$, $\boldsymbol{mFARVAT_B^{Het}}$, $\boldsymbol{mFARVAT_O^{Het}}$, $\boldsymbol{mFARVAT_B^{Hom}}$, $\boldsymbol{mFARVAT_S^{Hom}}$ and $\boldsymbol{mFARVAT_O^{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes (Q = 2 and Q = 5) with homogeneous and heterogeneous effects and different correlations (c = 0.5 and c = 0.8) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

Q	Type	c	Eff	*FARVAT*			$mFARVAT^{Het}$			$mFARVAT^{Hom}$		
				SKAT	Burden	SKAT-O	SKAT	Burden	SKAT-O	SKAT	Burden	SKAT-O
2	D	0.5	Het	0.038	0.000	0.028	0.069	0.000	0.048	0.020	0.000	0.016
			Hom	0.050	0.000	0.031	0.120	0.003	0.081	0.181	0.002	0.133
		0.8	Het	0.064	0.000	0.038	0.087	0.000	0.068	0.016	0.000	0.010
			Hom	0.052	0.003	0.039	0.100	0.007	0.082	0.127	0.004	0.101
	Q	0.5	Het	0.330	0.000	0.236	0.489	0.001	0.386	0.121	0.001	0.070
			Hom	0.335	0.001	0.240	0.481	0.003	0.405	0.657	0.005	0.566
		0.8	Het	0.341	0.001	0.246	0.431	0.000	0.327	0.103	0.000	0.064
			Hom	0.312	0.002	0.230	0.410	0.008	0.329	0.533	0.007	0.433

5	D	0.5	Het	0.067	0.001	0.038	0.409	0.001	0.320	0.043	0.000	0.024
			Hom	0.065	0.002	0.105	0.499	0.018	0.434	0.763	0.009	0.687
		0.8	Het	0.073	0.001	0.036	0.382	0.000	0.282	0.007	0.000	0.006
			Hom	0.060	0.000	0.044	0.381	0.009	0.305	0.557	0.003	0.445
	Q	0.5	Het	0.529	0.001	0.365	0.944	0.000	0.906	0.043	0.000	0.024
			Hom	0.472	0.001	0.333	0.883	0.018	0.836	0.983	0.012	0.972
		0.8	Het	0.543	0.000	0.371	0.913	0.000	0.866	0.019	0.000	0.012
			Hom	0.411	0.001	0.277	0.817	0.008	0.744	0.918	0.005	0.875

The definition of the acronyms in Table 3.4: 1) *Q*: the number of phenotypes; 2) Type: the type of phenotypes, D – dichotomous, Q - quantitative; 4) *c*: the correlation among the phenotypes; 5) Eff: the underlying genetic effects architecture, Hom – homogeneous effects, Het – heterogeneous effects; 6) SKAT: the sequence kernel association test; 7) Burden: the burden test; 8) SKAT-O: the optimal SKAT.

Table 3.5 Empirical power estimates when half rare variants are causal and 100% of them are deleterious. Empirical power of $mFARVAT_S^{Het}$, $mFARVAT_B^{Het}$, $mFARVAT_O^{Het}$, $mFARVAT_B^{Hom}$, $mFARVAT_S^{Hom}$ and $mFARVAT_O^{Hom}$ was calculated for dichotomous and quantitative multiple phenotypes (Q = 2 and Q = 5) with homogeneous and heterogeneous effects and different correlation (c = 0.5 and c = 0.8) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

Q	Type	c	Eff	*FARVAT*			$mFARVAT^{Het}$			$mFARVAT^{Hom}$		
				SKAT	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.207	0.273	0.427	0.340	0.523	0.663	0.219	0.488	0.569
			Hom	0.259	0.298	0.484	0.403	0.578	0.712	0.502	0.536	0.729
		0.8	Het	0.425	0.634	0.788	0.572	0.831	0.910	0.379	0.826	0.863
			Hom	0.488	0.684	0.834	0.678	0.857	0.937	0.812	0.851	0.948
	Q	0.5	Het	0.178	0.254	0.419	0.285	0.422	0.575	0.139	0.382	0.461
			Hom	0.248	0.283	0.462	0.390	0.497	0.647	0.420	0.460	0.631
		0.8	Het	0.400	0.602	0.754	0.521	0.759	0.877	0.251	0.752	0.794
			Hom	0.434	0.646	0.767	0.562	0.781	0.883	0.645	0.768	0.901

5	D	0.5	Het	0.294	0.427	0.630	0.839	0.921	0.977	0.417	0.908	0.910
			Hom	0.375	0.512	0.722	0.886	0.963	0.990	0.952	0.951	0.995
		0.8	Het	0.609	0.807	0.920	0.974	0.997	0.999	0.645	0.997	0.997
			Hom	0.665	0.845	0.944	0.977	0.999	1.000	0.999	0.998	1.000
	Q	0.5	Het	0.266	0.383	0.582	0.729	0.817	0.917	0.276	0.812	0.817
			Hom	0.328	0.464	0.651	0.773	0.867	0.947	0.854	0.835	0.960
		0.8	Het	0.595	0.759	0.901	0.900	0.961	0.991	0.405	0.955	0.956
			Hom	0.631	0.782	0.911	0.919	0.973	0.996	0.963	0.969	0.998

The definition of the acronyms in Table 3.5: 1) *Q*: the number of phenotypes; 2) Type: the type of phenotypes, D – dichotomous, Q - quantitative; 4) *c*: the correlation among the phenotypes; 5) Eff: the underlying genetic effects architecture, Hom – homogeneous effects, Het – heterogeneous effects; 6) SKAT: the sequence kernel association test; 7) Burden: the burden test; 8) SKAT-O: the optimal SKAT.

Table 3.6 Empirical power estimates when half rare variants are causal and 80% of them are deleterious. Empirical power of $\boldsymbol{mFARVAT_S^{Het}}$, $\boldsymbol{mFARVAT_B^{Het}}$, $\boldsymbol{mFARVAT_O^{Het}}$, $\boldsymbol{mFARVAT_B^{Hom}}$, $\boldsymbol{mFARVAT_S^{Hom}}$ and $\boldsymbol{mFARVAT_O^{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes (Q = 2 and Q = 5) with homogeneous and heterogeneous effects and different correlation (c = 0.5 and c = 0.8) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

Q	Type	c	Eff	*FARVAT*			$mFARVAT^{Het}$			$mFARVAT^{Hom}$		
				SKAT	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.128	0.043	0.174	0.215	0.105	0.306	0.098	0.091	0.182
			Hom	0.167	0.055	0.217	0.295	0.165	0.392	0.369	0.132	0.419
		0.8	Het	0.392	0.114	0.453	0.565	0.215	0.640	0.219	0.200	0.384
			Hom	0.413	0.138	0.490	0.601	0.301	0.705	0.751	0.272	0.793
	Q	0.5	Het	0.112	0.045	0.164	0.169	0.079	0.238	0.072	0.062	0.135
			Hom	0.159	0.052	0.203	0.240	0.133	0.327	0.301	0.112	0.348
		0.8	Het	0.375	0.112	0.410	0.469	0.152	0.526	0.137	0.135	0.267
			Hom	0.391	0.145	0.477	0.514	0.245	0.611	0.625	0.223	0.672

5	D	0.5	Het	0.184	0.059	0.245	0.703	0.317	0.769	0.118	0.288	0.363
			Hom	0.254	0.108	0.345	0.773	0.518	0.848	0.907	0.458	0.913
		0.8	Het	0.581	0.152	0.604	0.968	0.469	0.975	0.209	0.452	0.568
			Hom	0.612	0.267	0.696	0.953	0.690	0.977	0.996	0.662	0.997
	Q	0.5	Het	0.237	0.049	0.194	0.612	0.211	0.651	0.062	0.187	0.234
			Hom	0.237	0.090	0.311	0.669	0.353	0.736	0.779	0.292	0.789
		0.8	Het	0.581	0.135	0.568	0.912	0.321	0.927	0.094	0.305	0.352
			Hom	0.573	0.197	0.631	0.875	0.512	0.911	0.943	0.459	0.953

The definition of the acronyms in Table 3.6: 1) *Q*: the number of phenotypes; 2) Type: the type of phenotypes, D – dichotomous, Q - quantitative; 4) *c*: the correlation among the phenotypes; 5) Eff: the underlying genetic effects architecture, Hom – homogeneous effects, Het – heterogeneous effects; 6) SKAT: the sequence kernel association test; 7) Burden: the burden test; 8) SKAT-O: the optimal SKAT.

Table 3.7 Empirical power estimates when half rare variants are causal and 50% of them are deleterious. Empirical power of $\boldsymbol{mFARVAT_S^{Het}}$, $\boldsymbol{mFARVAT_B^{Het}}$, $\boldsymbol{mFARVAT_O^{Het}}$, $\boldsymbol{mFARVAT_B^{Hom}}$, $\boldsymbol{mFARVAT_S^{Hom}}$ and $\boldsymbol{mFARVAT_O^{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes (Q = 2 and Q = 5) with homogeneous and heterogeneous effects and different correlation (c = 0.5 and c = 0.8) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

Q	Type	c	Eff	*FARVAT*			$mFARVAT^{Het}$			$mFARVAT^{Hom}$		
				SKAT	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.041	0.000	0.024	0.094	0.000	0.064	0.024	0.000	0.011
			Hom	0.054	0.001	0.036	0.148	0.004	0.113	0.196	0.001	0.139
		0.8	Het	0.343	0.001	0.239	0.500	0.001	0.403	0.127	0.000	0.090
			Hom	0.341	0.004	0.246	0.500	0.004	0.421	0.677	0.004	0.568
	Q	0.5	Het	0.050	0.000	0.036	0.071	0.001	0.050	0.020	0.000	0.012
			Hom	0.062	0.001	0.037	0.118	0.000	0.088	0.139	0.001	0.102
		0.8	Het	0.352	0.001	0.253	0.426	0.000	0.327	0.087	0.000	0.062
			Hom	0.333	0.001	0.232	0.430	0.002	0.356	0.554	0.002	0.448

5	D	0.5	Het	0.079	0.000	0.053	0.420	0.001	0.318	0.017	0.000	0.018
			Hom	0.130	0.000	0.086	0.532	0.014	0.468	0.771	0.009	0.703
		0.8	Het	0.535	0.003	0.367	0.940	0.001	0.891	0.046	0.000	0.031
			Hom	0.508	0.000	0.364	0.900	0.014	0.848	0.977	0.007	0.963
	Q	0.5	Het	0.075	0.010	0.042	0.363	0.000	0.264	0.002	0.000	0.002
			Hom	0.115	0.003	0.075	0.449	0.012	0.389	0.594	0.009	0.505
		0.8	Het	0.562	0.001	0.377	0.901	0.000	0.841	0.017	0.000	0.010
			Hom	0.466	0.003	0.338	0.815	0.014	0.759	0.914	0.010	0.874

The definition of the acronyms in Table 3.7: 1) *Q*: the number of phenotypes; 2) Type: the type of phenotypes, D – dichotomous, Q - quantitative; 4) *c*: the correlation among the phenotypes; 5) Eff: the underlying genetic effects architecture, Hom – homogeneous effects, Het – heterogeneous effects; 6) SKAT: the sequence kernel association test; 7) Burden: the burden test; 8) SKAT-O: the optimal SKAT.

I also compared the performance of *mFARVAT*^{Het} and *mFARVAT*^{Hom} using simulated data. Tables 3.2-3.7 show that if the effects of each rare variant on phenotypes are heterogeneous, *mFARVAT*^{Het} performs better than *mFARAVAT*^{Hom}, and *vice versa*. In addition, when the effects of causal variants go in different directions, as in cases where some variants are deleterious while others are protective, the gap between the power of *mFARVAT*^{Het} and *mFARAVAT*^{Hom} is larger than in a scenario where such effects are in the same direction. Interestingly, for each method the statistical power difference between 100% and 50% deleterious causal variants seems to be larger for family-based samples than that for population-based designs (Lee et al. 2012).

Results for dichotomous phenotypes tend to be similar to those for quantitative phenotypes, although statistical power for the former is usually smaller. This difference may be explained by the fact that dichotomous phenotypes were transformed from quantitative phenotypes. Moreover, overall the power is seen to be inversely related to correlations among phenotypes. There is some power loss when c is increased from 0.5 to 0.8. Notably, when more phenotypes are included in the analysis, *mFARVAT* performs more effectively.

Last, I compared the proposed method with univariate analyses using *FARVAT* (Choi et al. 2014). The minimum p-value adjusted by Bonferroni correction was selected to calculate the power of univariate analyses. I considered two scenarios: multiple phenotypes are associated with variants and

only a single phenotype is associated with variants. Results in Tables 3.2-3.7 show that for the former scenario multivariate rare variant analyses perform better than univariate analyses. For the latter scenario, univariate rare variant analyses outperform multivariate analyses (see Table 3.8).

Table 3.8 Empirical power estimates when only one phenotype is associated with a region to test. Taking the proportion of deleterious and protective variants among 60 causal rare variants to be 100/0, 80/20, and 50/50, and only one phenotype was associated with the rare variants, the empirical power of $\boldsymbol{mFARVAT_S^{Het}}$, $\boldsymbol{mFARVAT_B^{Het}}$, $\boldsymbol{mFARVAT_O^{Het}}$, $\boldsymbol{mFARVAT_B^{Hom}}$, $\boldsymbol{mFARVAT_S^{Hom}}$ and $\boldsymbol{mFARVAT_O^{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes (Q = 2 and Q = 5) with correlation (c = 0.5) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

Q	Type	+/-	*FARVAT*			*mFARVAT*^{Het}			*mFARVAT*^{Hom}		
			SKAT	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	100/0	0.130	0.520	0.540	0.020	0.070	0.090	0.010	0.060	0.060
		80/20	0.020	0.120	0.150	0.000	0.010	0.020	0.000	0.020	0.020
		50/50	0.040	0.010	0.030	0.000	0.000	0.000	0.000	0.000	0.000
	Q	100/0	0.190	0.940	0.950	0.000	0.170	0.150	0.000	0.160	0.100
		80/20	0.110	0.310	0.450	0.010	0.050	0.060	0.000	0.050	0.060
		50/50	0.170	0.000	0.110	0.020	0.000	0.010	0.000	0.000	0.000
	D	100/0	0.050	0.420	0.450	0.000	0.000	0.000	0.010	0.000	0.000
		80/20	0.040	0.070	0.120	0.010	0.000	0.010	0.000	0.000	0.000
		50/50	0.020	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.000

5	Q	100/0	0.140	0.800	0.790	0.000	0.000	0.000	0.000	0.000	0.000
		80/20	0.110	0.190	0.370	0.000	0.000	0.000	0.000	0.000	0.000
		50/50	0.090	0.000	0.110	0.000	0.000	0.000	0.000	0.000	0.000

The definition of the acronyms in Table 3.8: 1) *Q*: the number of phenotypes; 2) Type: the type of phenotypes, D – dichotomous, Q - quantitative; 4) +/-: the number of variants with positive/negative effect; 5) SKAT: the sequence kernel association test; 6) Burden: the burden test; 7) SKAT-O: the optimal SKAT.

3.4 Application to COPD data

I applied *mFARVAT* to whole-exome sequencing data from the Boston Early-onset COPD Study (EOCOPD) (Silverman et al. 1998). The EOCOPD data are derived from an extended pedigree-based design. Probands were selected by age ≤53 years old, prebronchodilator forced expiratory volume in one second (FEV_1) of ≤40%, physician-diagnosed COPD, and without severe alpha-1 antitrypsin deficiency. All first-degree relatives, older second-degree relatives, and additional affected family members were enrolled. 49 pedigrees with at least 2 affected family members were selected for WES. Sequencing was performed at the University of Washington (Seattle, WA) Center for Mendelian Genomics, using Nimblegen V2 capture (Roche NimbleGen, Inc., Madison, WI) and the Illumina platform (Illumina, Inc., San Diego, CA). Quality control was performed using PLINK (Purcell et al. 2007), vcfTools (Danecek et al. 2011), and PLINK/SEQ at Brigham and Women's Hospital. Quality control included Mendelian error rates (< 1%), HWE ($p > 10^{-8}$), and average sequencing depth (> 12). Relatedness of individuals was evaluated by comparing KCM and GRM. Heterozygous/homozygous genotype ratio, Mendelian errors, proportion of variants in dbSNP and proportion of non-synonymous variants were used to identify outliers. After additionally filtering out samples with missing phenotypes or covariates, 254 samples from 49 families were obtained. The descriptive details of the EOCOPD data are provided in Table 3.9.

Table 3.9 The description of early-onset chronic obstructive pulmonary disease (EOCOPD) data. This description includes the range of age, and the number of individuals, families, females/males, cases/controls, variants, rare variants (MAF <5%) and genes.

Description	EOCOPD
Age	[21, 87]
Sample size	254
Families	49
F/M	172/82
Cases/controls	132/122
Variants	124,288
Rare variants	88,373
Genes	8,126

The definition of the acronyms in Table 3.9: 1) Families: the number of families; 2) F/M: the number of females and males; 3) Cases/controls: the number of cases and controls; 4) Variants: the number of variants; 5) Rare variants: the number of rare variants; 6) Genes: the number of genes.

I considered five COPD-related phenotypes: forced expiratory volume in one second pre-bronchodilator (FEVPRE); forced vital capacity post-bronchodilator (FVCPST); forced expiratory flow 25-75% pre-bronchodilator (DPRF2575); FEVPRE divided by FVCPRE (RATIO); and DPRF2575 divided by FVCPRE (F2575RAT). Sex, age, height, and pack-years of cigarette smoking were utilized to estimate BLUP offsets. It should be noted that genotypes were not used to estimate offsets. The correlation structure of the phenotypes is shown in Table 3.10.

Table 3.10 Correlation structure of the five chronic obstructive pulmonary disease (COPD) related phenotypes.

	FEVPRE	FVCPST	DPRF2575	F2575RAT	RATIO
FEVPRE	1	0.907	0.919	0.770	0.825
FVCPST	0.907	1	0.734	0.497	0.569
DPRF2575	0.919	0.734	1	0.919	0.829
F2575RAT	0.770	0.497	0.919	1	0.898
RATIO	0.825	0.569	0.829	0.898	1

The definition of the acronyms in Table 3.10: 1) FEVPRE: forced expiratory volume in one second pre-bronchodilator; 2) FVCPST: forced vital capacity post-bronchodilator; 4) DPRF2575: forced expiratory flow 25-75% pre-bronchodilator; 5) F2575RAT: DPRF2575/FVCPRE.

I assumed that variants with MAFs less than 5% were rare, and considered only genes with at least two rare variants and a MAC of at least four. As a result, 8,126 genes and 88,373 rare variants were analyzed. Our statistic requires the correlation matrix between individuals to obtain $\mathbf{\Phi}$. If there exists population substructure, GRM should be utilized for $\mathbf{\Phi}$ and otherwise KCM is adequate. I found no significant population substructure, and KCM was used for $\mathbf{\Phi}$. The Bonferroni-corrected 0.05 genome-wide significance level is 6.15E-6. QQ plots in Figures 3.6 show the statistical validity of our analysis. Manhattan plots are shown in Figure 3.7. The top 10 most significant results from $mFARVAT^{Het}$ and $mFARVAT^{Hom}$ are shown in Table 3.11. I could not find any genome-wide significant results with association analysis of multiple phenotypes. The most significant result was found for *KRTAP5-9* on chromosome 11, with $mFARVAT^{Het}$ (p-value = 1.00×10^{-4}), but the p-value for *KRTAP5-9* from $mFARVAT^{Hom}$ is 2.72×10^{-4}. The smaller p-value of $mFARVAT^{Het}$ may indicate that effect of each rare variant on the multiple phenotypes is heterogeneous.

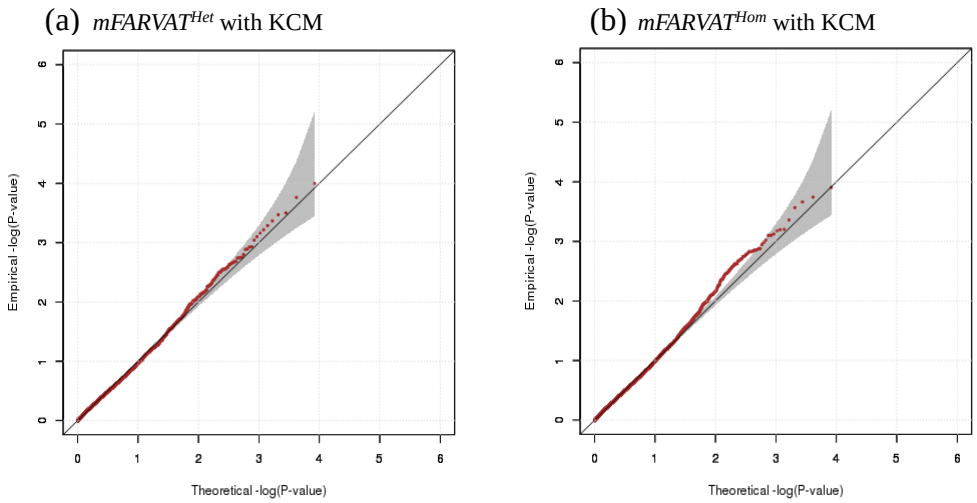


Figure 3.6 QQ plots of *mFARVAT* with the COPD data. *mFARVAT* was applied to the five COPD-related phenotypes. QQ-plots in (a) and (b) were from heterogeneous and homogeneous *mFARVAT*, respectively. KCM was utilized as Φ.

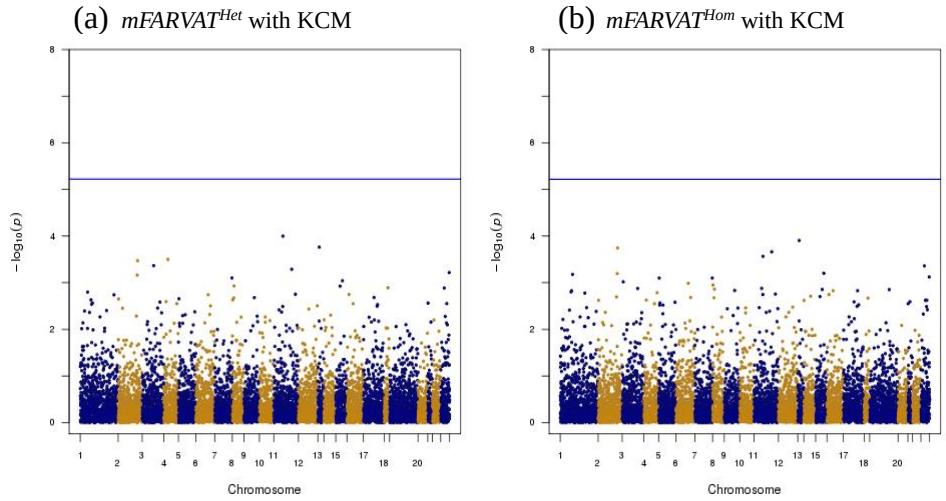


Figure 3.7 Manhattan plots of *mFARVAT* with the COPD data. *mFARVAT* was applied to the five COPD-related phenotypes. Manhattan plots in (a) and (b) are from heterogeneous and homogeneous *mFARVAT*, respectively. KCM was utilized as Φ.

Table 3.11 *mFARVAT* analysis of the COPD-releted phenotypes. Genes are the top 10 most significant results from $\boldsymbol{mFARVAT_O^{Het}}$ and $\boldsymbol{mFARVAT_O^{Hom}}$.

Method	Chr	Gene	MAC	# variants	P-value
Het	11	*KRTAP5-9*	21	3	1.00×10^{-04}
	13	*DIAPH3*	40	7	1.73×10^{-04}
	4	*ENAM*	82	9	3.16×10^{-04}
	2	*SLC8A1*	5	3	3.38×10^{-04}
	3	*MFI2*	32	5	4.30×10^{-04}
	11	*PLEKHA7*	20	9	5.16×10^{-04}
	2	*SLC19A3*	11	4	6.88×10^{-04}
	7	*ZNF736*	8	2	7.94×10^{-04}
	15	*MGA*	49	11	9.08×10^{-04}
	8	*CA1*	7	2	1.18×10^{-03}
Hom	13	*DIAPH3*	40	7	1.25×10^{-04}
	2	*SLC8A1*	5	3	1.80×10^{-04}
	11	*PLEKHA7*	20	9	2.18×10^{-04}
	11	*KRTAP5-9*	21	3	2.72×10^{-04}
	15	*POLG*	58	8	6.28×10^{-04}
	2	*SLC19A3*	11	4	6.37×10^{-04}
	1	*ETV3L*	31	5	6.63×10^{-04}
	7	*ZNF736*	8	2	7.94×10^{-04}
	5	*AFAP1L1*	20	3	7.95×10^{-04}
	3	*ANO10*	32	3	9.57×10^{-04}

The definition of the acronyms in Table 3.11: 1) Chr: chromosome; 2) MAC: minor allele count; 3) # variants: the number of rare variants; 4) Het: heterogeneous *mFARVAT*; 5) Hom: homogeneous *mFARVAT*.

3.5 Discussion

Extended families have complex correlation structure and association analyses using extended families are very complicated, in particular for dichotomous phenotypes. For instance, the unbalanced nature of family-based samples can lead to inflation or deflation of sandwich estimators for the variance-covariance matrix, and results from generalized estimating equation can be invalid (Aaij et al. 2013). An alternative approach is to use a generalized linear mixed model. However, calculating maximum likelihood estimators requires numerical integration, which is computationally very intensive, and approximations to avoid this can introduce serious bias (Gilmour et al. 1985, Schall 1991). Therefore in spite of the efficiency of extended families for rare variant association analysis, few methods have been suggested for family-based association analyses. In this chapter, I propose a new method of family-based analysis of rare variants associated with dichotomous phenotypes, quantitative phenotypes, or both. The proposed method enables multivariate analyses of extended families to detect rare variants. Extensive simulation studies show that *mFARVAT* works well for dichotomous and quantitative phenotypes. Our method is computationally efficient and association analyses at the genome-wide scale are computationally feasible for extended families. In our analyses, an Intel (R) Xeon (R) E5-2620 0 CPU at 2.00GHz, with a single node and 80 gigabyte memory, required six minutes to analyze the real data on two

phenotypes. *mFARVAT* is implemented in C++ and freely downloadable from http://healthstat.snu.ac.kr/software/mfarvat.

However, in spite of the analytical flexibility and efficiency of the method, some limitations still remain. First, GRM should ideally be used as the correlation matrix $\mathbf{\Phi}$ to provide robustness against population substructure; however, proper estimation of GRM requires large-scale common variants. In the absence of such data, the transmission disequilibrium test (Laird et al. 2000) is a unique alternative. Second, the proposed statistics are for retrospective designs and power loss is expected if samples are prospectively gathered. It has been shown that appropriate choice of offset minimizes power loss in certain scenarios but further investigation is still necessary. Third, *mFARVAT* cannot be used directly to analyze X-linked variants. The distribution of X-linked genetic variants in male is different from that in female, and thus different statistics for males and females are required. This issue will be investigated in my future work. Forth, homogeneous model and heterogeneous model are powerful when the real genetic model satisfies their assumptions. Specifically, if the effects of a variant among different phenotypes are in the same direction, homogeneous *mFARVAT* is more powerful, otherwise, heterogeneous *mFARVAT* performs better. However, the underlying genetic architecture is usually unknown. It would be more practical if we can propose a combined omnibus method, which can combine homogeneous and heterogeneous statistics and can be more robust for various genetic models. Similar to SKAT-O, the combined omnibus method can be derived as:

$$MS_c = cMS_{\rho_1}^{Hom} + (1-c)MS_{\rho_2}^{Het},$$

where $MS_{\rho_1}^{Hom}$ and $MS_{\rho_2}^{Het}$ are the homogeneous and heterogeneous *mFARVAT* SKAT-O statistics; $0 \leq \rho_1,\ \rho_2 \leq 1$ are values that make both statistics reach the minimum p-values, respectively. ρ_1, ρ_2 can be interpreted as the pairwise correlation among the genetics effects of different variants; $0 \leq c \leq 1$ can be interpreted as a pairwise correlation among the genetic effects of different phenotypes. Similarly, we can calculate the p-values of MS_c, pMS_c, with a grid of *c* and choose the one by the minimum p-value, P_{min}^c. For example,

$$P_{min}^c = \min\{pMS_0, pMS_{0.1^2}, \ldots, pMS_{0.5^2}, pMS_1\}.$$

P_{min}^c is the actual statistic of this combined omnibus method. To calculate the p-value of the omnibus statistic, the traditional SKAT-O derives the linear combination into a three independent terms and calculate p-value with one-dimensional numerical integration. However, this approach is not feasible here since $MS_{\rho_1}^{Hom}$ and $MS_{\rho_2}^{Het}$ contains different scores. Permutation is not applicable neither to family-based designs. Gene-dropping algorithm is a promising solution for this issue and will be considered in my future work.

Over the last decade, we have recognized that a substantial amount of unidentified genetic risk exists, and much effort has been expended to investigate this risk. Our methods provide an efficient strategy to analyze rare

variant associations in family-based samples, and it may increase understanding of heritable diseases.

This chapter was accepted in *Frontiers in Genetics*
as a partial fulfillment of Longfei Wang's PhD program.

Chapter 4

Family-based Rare Variant Association Test for Meta-analysis

4.1 Introduction

In this chapter, I proposed a new meta-analysis method for family-based, population-based, and case-control rare variant association tests, *metaFARVAT*. *metaFARVAT* generates a quasi-likelihood score for each variant and combines them to generate burden, VT, SKAT, and SKAT-O statistics. *metaFARVAT* can assume homogeneous or heterogeneous effects of variants among different studies and can be applied to both quantitative and dichotomous phenotypes. I evaluated the statistical validity of *metaFARVAT* using simulated data and

compared its estimated power with those of RAREMETAL and seqMeta under various scenarios. Furthermore, *metaFARVAT* was applied to identify rare variants for COPD using whole-exome sequencing (WES) data from family-based samples from the Boston Early-Onset COPD Study (EOCOPD) and case-control samples from the COPDGene study.

4.2 Methods

4.2.1 Notation

I assume that there are *K* studies available and that each study is of either a population-based, case-control, or family-based design. It is assumed that N_k subjects are available in study *k*. I assume that there are *M* rare variants in a gene, and the MAC of variant *m* for subject *j* in study *k* is coded by g_{jmk}. Traits can be either quantitative or dichotomous, and y_{jk} indicates a phenotype of subject *j* in study *k*. Their vectors are denoted by

$$\mathbf{G}_k^m = \begin{bmatrix} x_{1mk} \\ \vdots \\ x_{N_k mk} \end{bmatrix}, \mathbf{G}_k = \left(\mathbf{G}_k^1, \cdots, \mathbf{G}_k^M\right), \mathbf{Y}_k = \begin{bmatrix} y_{1k} \\ \vdots \\ y_{N_k k} \end{bmatrix}.$$

In some cases, rare variants may be observed only in a subset of studies. If variant *m* is missing or monomorphic in study *k*, I assume that $\mathbf{G}_k^m$ is $\mathbf{0}$, and its variance and covariance with $\mathbf{G}_k^{m'} (m \neq m')$ are 0. If variant *m* is missing for all studies, then it should be removed from the analysis.

Parental genotypes are transmitted to offspring under Mendelian transmission, and thus our test statistics consider the genetic correlation between family members. The genetic variance-covariance matrix among family members can be specified by a kinship coefficient matrix, $\mathbf{\Phi}_k$. Under the presence of population substructure, the genetic relationship matrix (GRM) can be estimated with large-scale genotyping data and should alternatively replace $\mathbf{\Phi}_k$ (Thornton et al. 2012).

Last, meta-analysis of rare variant association analyses with multiple studies requires two different types of weights. First, when multiple studies are combined, each study has different features, such as sample size and disease diagnosis, and such differences can be handled with an *a priori* specified weight for each study. I assume that the statistics for study *k* are weighted by v_k, and their *K*×*K* dimensional diagonal matrix is denoted by $\mathbf{W}_B$. Second, rare variants have different gene annotations, genomic coordinates, and functional characterization, and various annotation tools have been proposed to choose important features based on their biological properties. I denote the weight for rare variant *m* by w_m, and I let $\mathbf{W}_W$ be their *M* ×*M* dimensional diagonal matrix.

4.2.2 Choices of Offset

I introduce the offset μ_{jk} for subject *j* at study *k* to improve the efficiency of the proposed score test (Lange et al. 2002). I set

$$\boldsymbol{\mu}_k = \begin{bmatrix} \mu_{1k} \\ \vdots \\ \mu_{N_k k} \end{bmatrix}, \boldsymbol{\mu} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K), \mathbf{T}_k = \mathbf{Y}_k - \boldsymbol{\mu}_k.$$

The most efficient choice of $\boldsymbol{\mu}$ may depend on the sampling scheme, and either the BLUP with covariates or the prevalence were shown to be the most efficient (Won and Lange 2013). If families are randomly selected, BLUP was shown to be the most efficient (Won and Lange 2013); otherwise, the prevalence is recommended for dichotomous phenotypes (Thornton and McPeek 2007, Won and Lange 2013). In this report, I focus on randomly selected families, and I

incorporate BLUP from the linear mixed model for **μ**. Under the null hypothesis, the linear mixed model (George and Elston 1987) for a quantitative phenotype is given by

$$\mathbf{Y} = \mathbf{X\alpha} + \mathbf{B} + \mathbf{E}, \mathbf{B} \sim MVN\left(0, \sigma_b^2 \mathbf{\Phi}\right) \; and \; \mathbf{E} \sim MVN(0, \sigma_e^2 \mathbf{I}_N),$$

where **X** is the covariate matrix and **α** is its regression coefficient vector; **B** and **E** indicate the polygenetic random effect and random error, respectively. Then, incorporation of BLUP as an offset gives

$$\mathbf{T} = \mathbf{Y} - \boldsymbol{\mu} = \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X^t H^{-1} X}\right)^{-1} \mathbf{X^t H^{-1}} - \hat{\sigma}_1^2 \mathbf{\Phi P}\right) \mathbf{Y},$$

where $\mathbf{H} = \hat{\sigma}_b^2 \mathbf{\Phi} + \hat{\sigma}_e^2 \mathbf{I}_N$, and $\mathbf{P} = \mathbf{H^{-1}} - \mathbf{H^{-1} X}\left(\mathbf{X^t H^{-1} X}\right)^{-1} \mathbf{X^t H^{-1}}$. For a dichotomous phenotype, use of the generalized linear mixed model might be considered an appropriate approach, but I estimated **T** in the same way as for quantitative phenotypes when individuals were randomly selected because of its superior statistical power (Won and Lange 2013).

4.2.3 Score for Quasi-likelihood

I let $\mathbf{1}w$ be a $w \times 1$ column vector, of which the elements are 1. The score based on quasi-likelihood for variant *m* in study *k* is defined by

$$u_{m,k} = \mathbf{T}_k^t \left(\mathbf{I}_{N_k} - \mathbf{1}_{N_k} \left(\mathbf{1}_{N_k}^t \mathbf{\Phi}_k^{-1} \mathbf{1}_{N_k}\right)^{-1} \mathbf{1}_{N_k}^t \mathbf{\Phi}_k^{-1}\right) \mathbf{G}_k^m.$$

If I denote the covariance between $g_{m,k}$ and $g_{m',k}$ by $\sigma_{mm',k}$, then $\mathrm{cov}\left(\mathbf{G}_k^m, \mathbf{G}_k^{m'}\right) = \sigma_{mm',k}\boldsymbol{\Phi}_k$, and $\sigma_{mm',k}$ is estimated by the empirical covariance. I let

$$\boldsymbol{\Sigma_k} = \begin{bmatrix} \sigma_{11,k} & \cdots & \sigma_{1M,k} \\ \vdots & \ddots & \vdots \\ \sigma_{M1,k} & \cdots & \sigma_{MM,k} \end{bmatrix}.$$

If I let $\mathbf{A}_k = \boldsymbol{\Phi}_k - \mathbf{1}_{N_k}\left(\mathbf{1}_{N_k}^t \boldsymbol{\Phi}_k^{-1} \mathbf{1}_{N_k}\right)^{-1} \mathbf{1}_{N_k}^t$, the variance-covariance matrix of $u_{m,k}$ (Choi et al. 2014) was shown to be

$$\mathrm{var}\begin{bmatrix} \mathbf{T}_k^t\left(\mathbf{G}_k^1 - \hat{E}\left(\mathbf{G}_k^1\right)\right) \\ \vdots \\ \mathbf{T}_k^t\left(\mathbf{G}_k^M - \hat{E}\left(\mathbf{G}_k^M\right)\right) \end{bmatrix} = (\mathbf{T}_k^t \mathbf{A}_k \mathbf{T}_k)\boldsymbol{\Sigma}_k.$$

The score vector of rare variants in study *k* can be defined by

$$\mathbf{U}_k = \frac{1}{\sqrt{\mathbf{T}_k^t \mathbf{A}_k \mathbf{T}_k}} \mathbf{T}_k^t \left(\mathbf{I}_{N_k} - \mathbf{1}_{N_k}\left(\mathbf{1}_{N_k}^t \boldsymbol{\Phi}_k^{-1} \mathbf{1}_{N_k}\right)^{-1} \mathbf{1}_{N_k}^t \boldsymbol{\Phi}_k^{-1}\right) \mathbf{G}_k.$$

The score statistic tests whether the coded genotypes are linearly independent from the phenotypes; for dichotomous phenotypes, it is equivalent to comparing the MAFs between cases and controls.

4.2.4 Homogeneous Model

The homogeneous model assumes that the effect sizes of each variant are expected to be in the same direction among different studies, and thus the proposed scores for each study can be collapsed across studies as follows:

$$\mathbf{U}^{Hom} \equiv \sum_k v_k \mathbf{U}_k^t\,,\ \boldsymbol{\Sigma}^{Hom} \equiv \mathrm{var}(\mathbf{U}^{Hom}) = \sum_k v_k^2 \boldsymbol{\Sigma}_k.$$

Here, I set v_k to be 1. However, the proposed statistics are sometimes unavailable, and the appropriate choice can vary according to the available information. For instance, if standardized test statistics and sample sizes are available, then the inverse function to the square root of the sample size can be utilized.

Rare variant association analysis can be categorized into burden and variance-component tests (Li and Leal 2008, Price et al. 2010, Neale et al. 2011, Wu et al. 2011). The burden test is known to be the most powerful if all rare variants have either deleterious or protective effects on disease; otherwise, the variance-component test is more efficient (Neale et al. 2011). If I let χ_1^2 be a chi-square distribution with a single degree of freedom, the burden test for a homogeneous model becomes

$$S_{burden}^{Hom} = \frac{(\mathbf{U}^{Hom})^t \mathbf{W}_W \mathbf{1}_M \mathbf{1}_M^t \mathbf{W}_W \mathbf{U}^{Hom}}{\mathbf{1}_M^t \mathbf{W}_W \boldsymbol{\Sigma}^{Hom} \mathbf{W}_W \mathbf{1}_M} \sim \chi_1^2 \text{ under } H_0.$$

Variance component tests use the collapsed squared scores (Neale et al. 2011, Wu et al. 2011) and can be expressed by

$$S_{SKAT}^{Hom} = (\mathbf{U}^{Hom})^t \mathbf{W}_W \mathbf{I}_M \mathbf{W}_W \mathbf{U}^{Hom}.$$

I denote eigenvalues for $(\boldsymbol{\Sigma}^{Hom})^{1/2} \mathbf{W}_W \mathbf{W}_W (\boldsymbol{\Sigma}^{Hom})^{1/2}$ by λ_m. If I let $\chi_{1,m}^2$ be an independent chi-square distribution with a single degree of freedom, the variance component test for the homogeneous model follows

$$S_{SKAT}^{Hom} \sim \sum_{m=1}^{M} \lambda_m \chi_{1,m}^2 \text{ under } H_0.$$

A balanced approach for both scenarios can be achieved by the SKAT-O type statistic (Lee et al. 2012). For a certain ρ between 0 and 1, I consider

$$(\mathbf{U}^{Hom})^t \mathbf{W}_W \left((1-\rho)\mathbf{I}_M + \rho \mathbf{1}_M \mathbf{1}_M^t\right) \mathbf{W}_W \mathbf{U}^{Hom}.$$

If I let its p-value be pS_ρ^{Hom}, the SKAT-O type statistic for $\rho_0 = 0 < \rho_1 < \cdots < \rho_L = 1$ is defined by

$$S_{SKATO}^{Hom} = p_{\min}^{Hom} = \min\{pS_0^{Hom}, pS_{0.01}^{Hom}, pS_{0.04}^{Hom}, pS_{0.09}^{Hom}, pS_{0.16}^{Hom}, pS_{0.25}^{Hom}, pS_1^{Hom}\}.$$

Its p-value can be calculated with the numerical algorithm for the FARVAT statistic (Choi et al. 2014).

Last, rare variant association analysis utilizes rare variants, but the definition of a rare variant is not clear. VT approaches are very useful in such scenarios. I assume that rare variants are sorted in ascending order of overall MAF. I let $\mathbf{1}_{(m)}$ be an M-dimensional column vector whose 1st, … , mth elements are 1 and the others are 0. If I let

$$U_{(m)}^{Hom} = \sum_{k=1}^{K} v_k \mathbf{1}_{(m)}^t \mathbf{W}_W \mathbf{U}_k^t = \mathbf{1}_{(m)}^t \mathbf{W}_W \mathbf{U}^{Hom},$$

then the covariance between $U_{(m)}^{Hom}$ and $U_{(m')}^{Hom}$ is

$$\mathbf{1}_{(m)}^t \mathbf{W}_W \boldsymbol{\Sigma}^{Hom} \mathbf{W}_W \mathbf{1}_{(m')}.$$

Therefore, I let

$$T_{(m)}^{Hom} = \frac{U_{(m)}^{Hom}}{\sqrt{\mathbf{1}_{(m)}^{t}\mathbf{W}_{W}\boldsymbol{\Sigma}^{Hom}\mathbf{W}_{W}\mathbf{1}_{(m)}}}.$$

If I denote the realization of $T_{(m)}^{Hom}$ by $t_{(m)}$ and let $t_{(|\max|)} = \max\{T_{(m)}^{Hom}\}$ (Spielman et al. 1993), the p-value for the VT method can be calculated by

$$1 - P\left(|T_{(1)}^{Hom}| > t_{(|\max|)}, \cdots, |T_{(M)}^{Hom}| > t_{(|\max|)}\right).$$

Here, $\left(T_{(1)}^{Hom}, \cdots, T_{(M)}^{Hom}\right)^{t}$ follows the multivariate normal distribution with mean 0 and the following variance-covariance matrix:

$$\boldsymbol{\Psi}^{\boldsymbol{Hom}} = \left(\boldsymbol{\Psi}_{\boldsymbol{mm'}}^{\boldsymbol{Hom}}\right)_{\boldsymbol{M}\times\boldsymbol{M'}}$$

where $\boldsymbol{\Psi}_{mm'}^{Hom} = \frac{\mathbf{1}_{(m)}^{t}\mathbf{W}_{w}\boldsymbol{\Sigma}^{Hom}\mathbf{W}_{w}\mathbf{1}_{(m')}}{\sqrt{\left(\mathbf{1}_{(m)}^{t}\mathbf{W}_{w}\boldsymbol{\Sigma}^{Hom}\mathbf{W}_{w}\mathbf{1}_{(m)}\right)\left(\mathbf{1}_{(m')}^{t}\mathbf{W}_{w}\boldsymbol{\Sigma}^{Hom}\mathbf{W}_{w}\mathbf{1}_{(m')}\right)}}$.

4.2.5 Heterogeneous Model

As in the homogeneous model, I propose burden and variance component tests for the heterogeneous model. The heterogeneous model assumes that the effects of specific variant *m* are heterogeneous among studies and follow an arbitrary distribution with mean 0 and variance τ_m. If I let $E\left(u_{m,k}\right) = \beta_{mk}$, the null hypothesis can be expressed by $\beta_{m1} = \cdots = \beta_{mK} = 0$, or simply $\tau_m = 0$, and I consider the following score vector and its variance matrix:

$$\mathbf{U}^{Het} \equiv (v_1\mathbf{U}_1 \quad \ldots \quad v_K\mathbf{U}_K)^{t}, \boldsymbol{\Sigma}^{Het} \equiv \mathrm{var}\left(\mathrm{vec}(\mathbf{U})\right) = \sum_k\left(v_k^2\boldsymbol{\Sigma}_k \otimes \boldsymbol{e}_{kk}\right),$$

where $\boldsymbol{e}_{kk}$ is a $K\times K$ dimensional matrix whose (k, k) element is 1 and the others are 0. Then, the burden test can be expressed as

$$S_{burden}^{Het} = \frac{\mathbf{U}^{Het}(\mathbf{W}_W\otimes\mathbf{W}_B)\mathbf{1}_{MK}\mathbf{1}_{MK}^t(\mathbf{W}_W\otimes\mathbf{W}_B)\mathbf{U}^{Het^t}}{\mathbf{1}_{MK}^t(\mathbf{W}_W\otimes\mathbf{W}_B)\boldsymbol{\Sigma}^{Het}(\mathbf{W}_W\otimes\mathbf{W}_B)\mathbf{1}_{MK}} \sim \chi_1^2 \text{ under } H_0.$$

I let

$$\mathbf{R}_c^{Het} = (1-c)\mathbf{I}_{MK} + c\mathbf{1}_{MK}\mathbf{1}_{MK}^t, S_c^{Het}$$

$$= \mathbf{U}^{Het}(\mathbf{W}_W\otimes\mathbf{W}_B)\mathbf{R}_c^{Het}(\mathbf{W}_W\otimes\mathbf{W}_B)\mathbf{U}^{Het^t},$$

and I let $(\lambda_1^c, \ldots, \lambda_{MK}^c)$ be the eigenvalues of

$$\sum_k(\boldsymbol{\Sigma}_k\otimes\boldsymbol{e}_{kk})\,(\mathbf{W}_W\otimes\mathbf{W}_B)\mathbf{I}_{MK}(\mathbf{W}_W\otimes\mathbf{W}_B)\sum_k(\boldsymbol{\Sigma}_k\otimes\boldsymbol{e}_{kk}).$$

Then, S_c^{Het} follows

$$S_c^{Het} \sim \sum_{l=1}^{MK}\lambda_l^c\chi_{1,l}^2 \text{ under } H_0.$$

Therefore, the variance component test is defined by

$$S_{SKAT}^{Het} = \mathbf{U}^{Het}(\mathbf{W}_W\otimes\mathbf{W}_B)\mathbf{I}_{MK}(\mathbf{W}_W\otimes\mathbf{W}_B)\mathbf{U}^{Het^t}$$

$$= S_0^{Het} \sim \sum_{l=1}^{MK}\lambda_l^0\chi_{1,l}^2 \text{ under } H_0.$$

If I denote the p-value for S_c^{Het} by pS_c^{Het}, the SKAT-O-type statistic is defined by

$$S_{SKATO}^{Het} = p_{\min}^{Het} = \min\{pS_0^{Het}, pS_{0.01}^{Het}, pS_{0.04}^{Het}, pS_{0.09}^{Het}, pS_{0.16}^{Het}, pS_{0.25}^{Het}, pS_1^{Het}\},$$

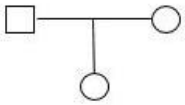
and its p-value is also obtained by the numerical algorithm for the *FARVAT* statistic (Choi et al. 2014)

4.3 Simulation study

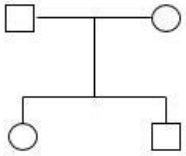
4.3.1 The simulation model

The performance of *metaFARVAT* was evaluated via extensive simulation studies. *metaFARVAT* can be applied to population-based and case-control designs by calculating GRM among samples. Therefore, I only focused on family-based designs in our simulation studies and considered unbalanced families consisting of trios, nuclear families, and extended families with 3 generations; the family structures that I considered are presented in Figure 4.1. The families for our simulations were randomly selected from these different family structures. To generate rare variants, 1,200 haplotypes with 50,000 base pairs were generated under a coalescent model using the software COSI (Schaffner et al. 2005).

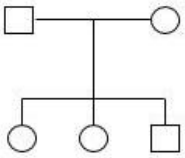
(a) Trio



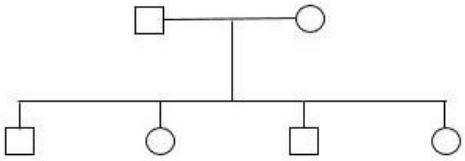
(b) Nuclear family with 4 members



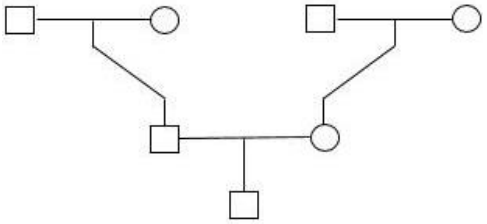
(c) Nuclear family with 5 members



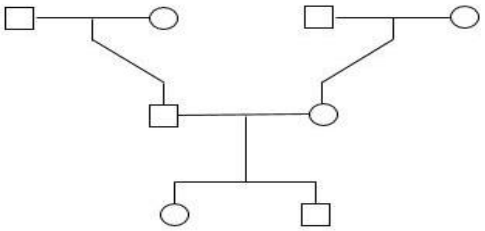
(d) Nuclear family with 6 members



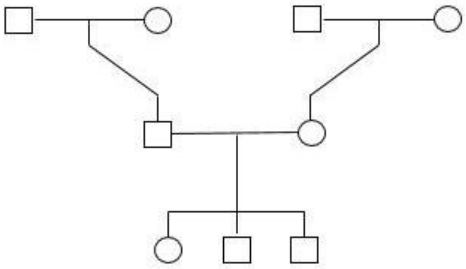
(e) Extended family with 7 members



(f) Extended family with 8 members



(g) Extended family with 9 members



(h) Extended family with 10 members

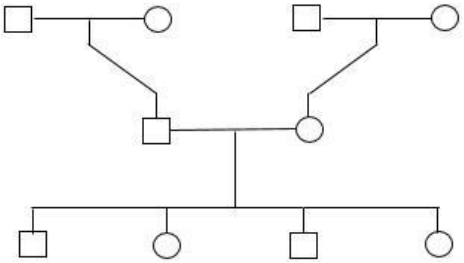


Figure 4.1 Family structures with different family members.

Each haplotype was generated by setting the mutation rate to 1.5×10^{-8}, and haplotypes were randomly chosen with replacement to build founder genotypes. I defined variants with MAFs < 0.01 as being rare, and 60 rare variants were randomly selected from their haplotypes. Then, non-founder haplotypes were chosen from their parents' haplotypes in Mendelian fashion under the assumption of no recombination.

Phenotypes were generated under the null and alternative hypotheses. Simulation of dichotomous phenotypes was performed using the liability threshold model. Once the quantitative phenotypes were generated, they were transformed into case-control status for dichotomous phenotypes. If quantitative phenotypes were larger than the threshold, they were considered affected and otherwise were considered unaffected. The threshold was chosen to preserve the assumed disease prevalence of 0.1. If the disease prevalence is misspecified, loss of statistical power is expected; however, it has been shown with simulation studies that the effect of misspecification is not very substantial (Won and Lange 2013). To allow for the ascertainment bias of dichotomous phenotypes in our simulation studies, I assumed that families with at least one affected subject were selected for analysis.

Quantitative phenotypes were defined by summing the phenotypic mean, polygenic effect, main genetic effect, and random error, and I assumed there was no environmental effect shared between family members. The phenotypic mean was denoted by α = 0.3. The polygenic effect for each founder was independently generated from $N(0, \sigma_g^2=1)$, and for non-founders, the average

of maternal and paternal polygenic effects was combined with values independently sampled from $N(0, 0.5\sigma_g^2)$. Random error was independently sampled from $N(0, \sigma_e^2=1)$. Therefore, the heritability of the simulated trait is 0.5. The genetic effect at variant m in study k was the product of β_{mk} and the number of disease susceptibility alleles. To evaluate the TIE estimates, β_{mk} was assumed to be 0. To evaluate the statistical power estimates, if I let h_a^2 be the proportion of variance explained by rare variants, β_{mk} values were iteratively sampled with a two-step approach. $\beta_{mk}^{(0)}$ were first sampled from $U(0,1)$. Then, if I let

$$v_k = \sqrt{\frac{(\sigma_g^2+\sigma_e^2)h_a^2}{(1-h_a^2)\sum_{m=1}^{M}[(\beta_{mk}^{(0)})^2 2p_m(1-p_m)]}},$$

β_{mk} values were sampled from the uniform distribution $U(0, v_k)$. This procedure was repeated until v_k converged. I assumed that $h_a^2 = 0.01$. β_{mk} was generated from heterogeneous or homogeneous scenarios. For homogeneous scenarios, I assumed that the effects of each rare variant were in the same direction in all studies. For heterogeneous scenarios, the signs (+/-) of β_{mk} values sampled from $U(0,v_k)$ were chosen randomly.

4.3.2 Evaluation of *metaFARVAT* with simulated data

To evaluate statistical validity, TIE estimates for both dichotomous and quantitative phenotypes were calculated at various significance levels using 20,000 replicates of 200 unbalanced families. For each replicate, I performed 3

different meta-analyses, including 3, 6, and 9 studies. Table 4.1 shows empirical TIE estimates for homogeneous *metaFARVAT* (*metaFARVAT*^{Hom}) and heterogeneous *metaFARVAT* (*metaFARVAT*^{Het}) at the 0.1, 0.01, 10^{-3}, and 10^{-4} significance levels with dichotomous phenotypes. Estimates of TIE rates were virtually equal to nominal significance levels. However, VT type *metaFARVAT*^{Hom} showed inflation, especially when there were 3 studies, and if the number of rare variants is small, it is not recommended. Quantile-quantile (QQ) plots in Figures 2–4 also show consistent results. Therefore, I conclude that the proposed *metaFARVAT*^{Hom} and *metaFARVAT*^{Het} are statistically valid.

Secondly, empirical power estimates for dichotomous phenotypes were calculated at the 2.5×10^{-6} significance level, showing the changes in power under different scenarios. Empirical power estimates were calculated with 2,000 replicates for 7 different statistics: burden, SKAT, SKAT-O, and VT type statistics for *metaFARVAT*^{Hom} and burden, SKAT, and SKAT-O type statistics for *metaFARVAT*^{Het}. Results are provided in Tables 4.2 and 4.3 for homogeneous and heterogeneous scenarios, respectively. In addition, I compared the proposed methods with two meta-analysis methods based on the use of p-values across studies: the minimum p-value method and Fisher's method. If I let p_k be the p-value from the kth study (k = 1,2,…,K), the minimum p-value and Fisher's method can be obtained by

$$\text{minP} = \min(p_k) \sim Beta(1, K), \ \text{Fisher} = -2\sum_{k=1}^{K} \ln p_k \sim \chi^2(df = 2K) \text{ under } H_0.$$

Table 4.1 Type I error estimates from simulation study with dichotomous phenotypes. The empirical type I error was estimated for the proposed methods with 20,000 replicates at the 0.1, 0.01, 10^{-3} and 10^{-4} significance levels for dichotomous phenotypes. I assumed that the number of rare variants is 60, and that their MAF <0.01. Both homogeneous (Hom) and heterogeneous (Het) models were considered.

Model	*K*	α	Dichotomous phenotype			
			Burden	SKAT	SKAT-O	VT
Hom	3	0.1	0.0960	0.0950	0.0953	0.1100
		0.01	0.0103	0.0099	0.0100	0.0116
		10^{-3}	0.0009	0.0012	0.0014	0.0017
		10^{-4}	0.0001	0.0001	0.0001	0.0004
	6	0.1	0.1002	0.0953	0.0957	0.1018
		0.01	0.0094	0.0085	0.0088	0.0106
		10^{-3}	0.0008	0.0009	0.0008	0.0011
		10^{-4}	0.0001	0.0000	0.0000	0.0001
	9	0.1	0.1000	0.1015	0.1025	0.1018
		0.01	0.0096	0.0098	0.0093	0.0110
		10^{-3}	0.0007	0.0009	0.0007	0.0015
		10^{-4}	0.0001	0.0000	0.0000	0.0001
Het	3	0.1	0.0987	0.1006	0.0981	--
		0.01	0.0100	0.0091	0.0094	--
		10^{-3}	0.0008	0.0008	0.0013	--
		10^{-4}	0.0001	0.0002	0.0002	--
	6	0.1	0.1036	0.0986	0.0985	--
		0.01	0.0094	0.0106	0.0105	--
		10^{-3}	0.0008	0.0014	0.0012	--
		10^{-4}	0.0001	0.0003	0.0002	--
	9	0.1	0.1041	0.1026	0.1046	--
		0.01	0.0107	0.0095	0.0107	--
		10^{-3}	0.0009	0.0011	0.0009	--
		10^{-4}	0.0001	0.0002	0.0001	--

The definition of the acronyms in Table 4.1: 1) *K*: the number of studies; 2) SKAT: the sequence kernel association test; 3) Burden: the burden test; 4) SKAT-O: the optimal SKAT; 5) VT: the variable threshold test; 6) α: significance level.

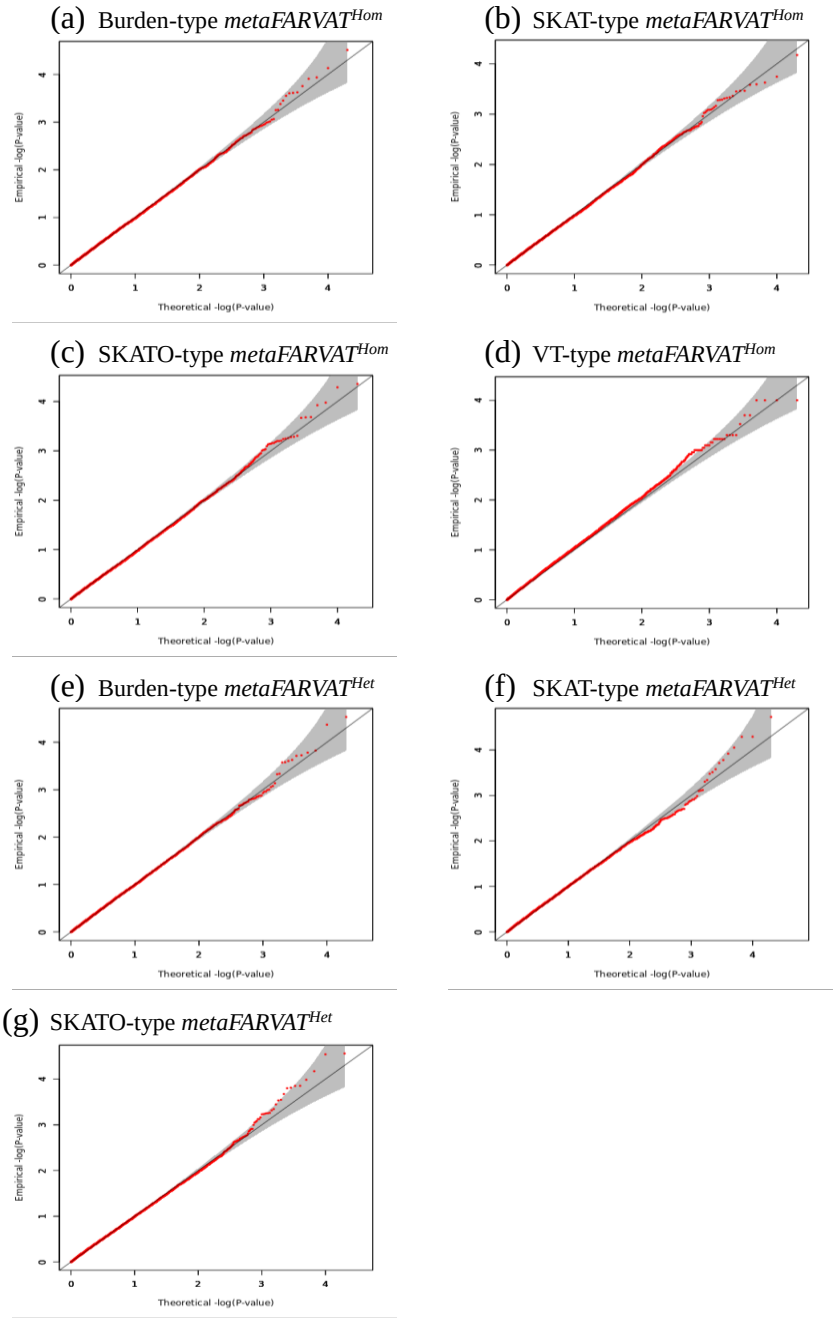
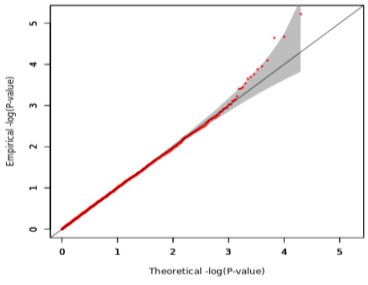
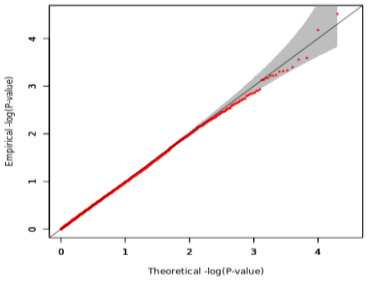


Figure 4.2 QQ plots for meta-analyses of dichotomous phenotype based on 3 studies. QQ plots were provided for results from the proposed methods under the null hypothesis. The empirical p-values were calculated under the null hypothesis with 20,000 replicates.

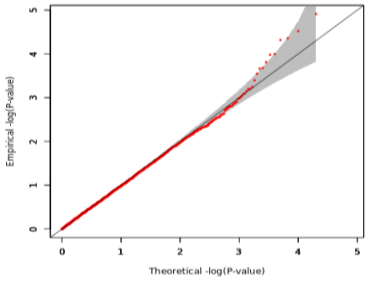
(a) Burden-type *metaFARVAT*^{Hom}



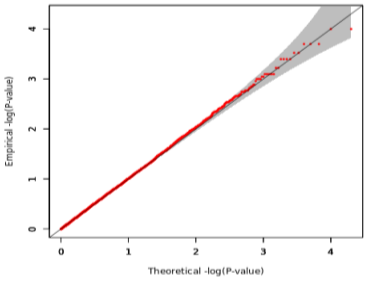
(b) SKAT-type *metaFARVAT*^{Hom}



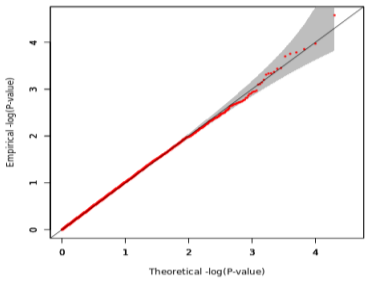
(c) SKATO-type *metaFARVAT*^{Hom}



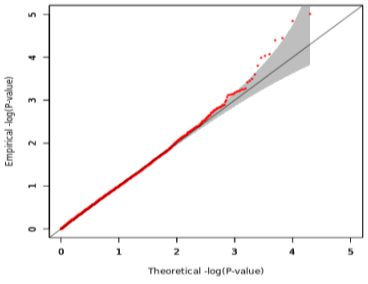
(d) VT-type *metaFARVAT*^{Hom}



(e) Burden-type *metaFARVAT*^{Het}



(f) SKAT-type *metaFARVAT*^{Het}



(g) SKATO-type *metaFARVAT*^{Het}

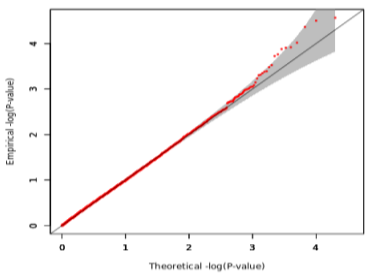
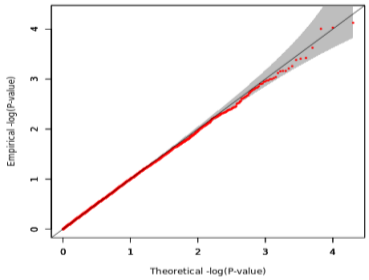
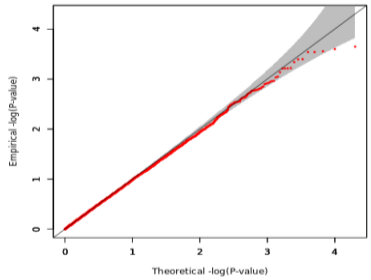


Figure 4.3 QQ plots for meta-analyses of dichotomous phenotype based on 6 studies. QQ plots were provided for results from the proposed methods under the null hypothesis. The empirical p-values were calculated under the null hypothesis with 20,000 replicates.

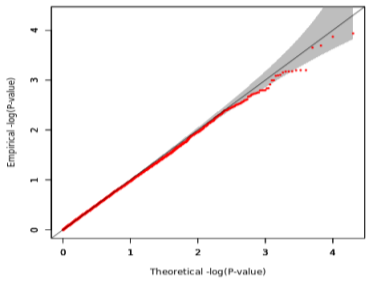
(a) Burden-type *metaFARVAT*^{Hom}



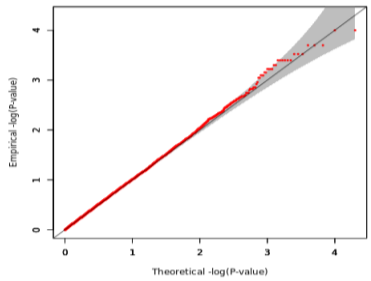
(b) SKAT-type *metaFARVAT*^{Hom}



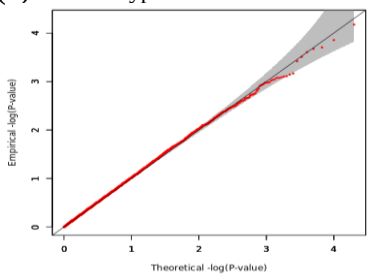
(c) SKATO-type *metaFARVAT*^{Hom}



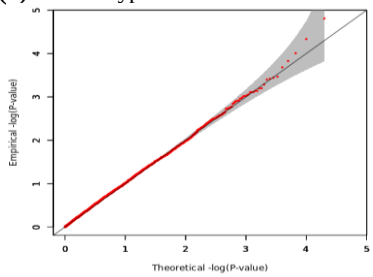
(d) VT-type *metaFARVAT*^{Hom}



(e) Burden-type *metaFARVAT*^{Het}



(f) SKAT-type *metaFARVAT*^{Het}



(g) SKATO-type *metaFARVAT*^{Het}

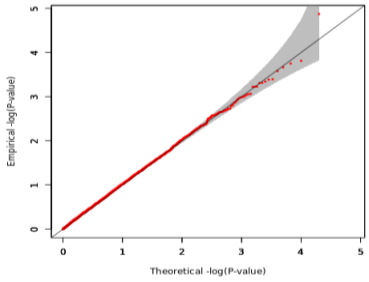


Figure 4.4 QQ plots for meta-analyses of dichotomous phenotype based on 9 studies. QQ plots were provided for results from the proposed methods under the null hypothesis. The empirical p-values were calculated under the null hypothesis with 20,000 replicates.

Table 4.2 Empirical power estimates for dichotomous phenotype for homogeneous variants among studies. Empirical power of burden, SKAT, SKAT-O and VT type of *metaFARVAT*^{Hom} and *metaFARVAT*^{Het} was calculated for dichotomous phenotypes with homogeneous effects at the 2.5×10^{-6} significance level.

+/-	Method	3 studies				6 studies				9 studies			
		SKAT	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**
60/0	Fisher	0.1990				0.6495				0.8940			
	minP	0.0315				0.0610				0.0715			
	Hom	0.0195	0.3590	0.3660	0.3915	0.1690	0.9265	0.9150	0.9240	0.4920	0.9975	0.9945	0.9965
	Het	0.0115	0.3390	0.4160	--	0.0750	0.9095	0.9330	--	0.1865	0.9930	0.9960	--
48/12	Fisher	0.0270				0.1060				0.2400			
	minP	0.0060				0.0070				0.0070			
	Hom	0.0105	0.0335	0.0670	0.0450	0.1105	0.2290	0.3720	0.2665	0.4000	0.5355	0.7565	0.5720
	Het	0.0045	0.0310	0.0720	--	0.0225	0.2080	0.3305	--	0.0760	0.4825	0.6325	--
30/30	Fisher	0.0000				0.0015				0.0035			
	minP	0.0000				0.0000				0.0000			
	Hom	0.0050	0.0000	0.0025	0.0010	0.0555	0.0000	0.0270	0.0000	0.2615	0.0000	0.1650	0.0065
	Het	0.0000	0.0000	0.0005	--	0.0020	0.0000	0.0015	--	0.0120	0.0000	0.0090	--
30/0	Fisher	0.0440				0.2090				0.4520			
	minP	0.0090				0.0170				0.0205			
	Hom	0.0140	0.0725	0.1145	0.0900	0.1790	0.4260	0.5760	0.4785	0.5515	0.7970	0.9125	0.8220
	Het	0.0070	0.0605	0.1290	--	0.0555	0.3905	0.5545	--	0.1410	0.7545	0.8590	--

24/6	Fisher	0.0075				0.0365				0.0895			
	minP	0.0020				0.0020				0.0015			
	Hom	0.0095	0.0045	0.0215	0.0085	0.1285	0.0465	0.1980	0.0610	0.4480	0.1440	0.5480	0.1765
	Het	0.0025	0.0035	0.0240	--	0.0225	0.0340	0.1215	--	0.0630	0.1105	0.2890	--
15/15	Fisher	0.0000				0.0030				0.0045			
	minP	0.0000				0.0010				0.0005			
	Hom	0.0020	0.0000	0.0005	0.0010	0.0550	0.0000	0.0270	0.0025	0.2700	0.0000	0.1650	0.0060
	Het	0.0000	0.0000	0.0000	--	0.0025	0.0000	0.0025	--	0.0090	0.0000	0.0030	--

The definition of the acronyms in Table 4.2: 1) +/-: the number of causal variants with positive and negative effect; 2) SKAT: the sequence kernel association test; 3) Burden: the burden test; 4) SKAT-O: the optimal SKAT; 5) VT: the variable threshold test; 6) Fisher: Fisher's method; 7) minP: the minimum p-value method; 8) Hom: homogeneous *metaFARVAT*; 9) Het: heterogeneous *metaFARVAT*.

Table 4.3 Empirical power estimates for dichotomous phenotype for heterogeneous variants among studies. Empirical power of burden, SKAT, SKAT-O and VT type of *metaFARVAT*^{Hom} and *metaFARVAT*^{Het} was calculated for dichotomous phenotypes with heterogeneous effects at the 2.5×10^{-6} significance level.

+/-	Method	3 studies				6 studies				9 studies			
		SKAT	Burden	SKAT-O	VT	SKAT	Burden	SKAT-O	VT	SKAT	Burden	SKAT-O	VT
48/12	Fisher	0.0240				0.1040				0.2235			
	minP	0.0065				0.0070				0.0105			
	Hom	0.0040	0.0340	0.0425	0.0460	0.0130	0.2170	0.2450	0.2555	0.0420	0.5200	0.5305	0.5680
	Het	0.0080	0.0325	0.0590	--	0.0270	0.1865	0.3180	--	0.0715	0.4555	0.6115	--
30/30	Fisher	0.0000				0.0030				0.0020			
	minP	0.0000				0.0000				0.0000			
	Hom	0.0005	0.0000	0.0000	0.0005	0.0005	0.0000	0.0005	0.0000	0.0015	0.0000	0.0015	0.0000
	Het	0.0005	0.0000	0.0005	--	0.0050	0.0000	0.0030	--	0.0090	0.0000	0.0070	--
30/0	Fisher	0.0460				0.2220				0.4690			
	minP	0.0115				0.0145				0.0160			
	Hom	0.0065	0.0670	0.0945	0.0880	0.0400	0.4385	0.4595	0.4730	0.1185	0.7880	0.7875	0.7980
	Het	0.0060	0.0570	0.1340	--	0.0510	0.3930	0.5580	--	0.1370	0.7425	0.8380	--
24/6	Fisher	0.0095				0.0325				0.0850			
	minP	0.0030				0.0035				0.0030			
	Hom	0.0020	0.0070	0.0115	0.0120	0.0045	0.0470	0.0680	0.0655	0.0125	0.1520	0.1875	0.1900
	Het	0.0040	0.0065	0.0270	--	0.0215	0.0335	0.1230	--	0.0590	0.1185	0.2825	--

15/15	Fisher	0.0010				0.0015				0.0020			
	minP	0.0005				0.0010				0.0005			
	Hom	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	0.0005
	Het	0.0015	0.0000	0.0015	--	0.0030	0.0000	0.0010	--	0.0095	0.0000	0.0060	--

The definition of the acronyms in Table 4.3: 1) +/-: the number of causal variants with positive and negative effect; 2) SKAT: the sequence kernel association test; 3) Burden: the burden test; 4) SKAT-O: the optimal SKAT; 5) VT: the variable threshold test; 6) Fisher: Fisher's method; 7) minP: the minimum p-value method; 8) Hom: homogeneous *metaFARVAT*; 9) Het: heterogeneous *metaFARVAT*.

According to our results, the minimum p-value approach usually performed the least efficiently, especially when there were equal numbers of protective and deleterious rare variants in the targeted gene. Moreover, the power of the minimum p-value approach was not much improved by including more studies in the meta-analysis. The Fisher approach always performed better than the minimum p-value approach but was less powerful than the *metaFARVAT* method, regardless of scenarios. Furthermore, the statistical power estimates of $metaFARVAT^{Het}$ were similar between the homogeneous and the heterogeneous scenarios. However, the statistical power estimates of $metaFARVAT^{Hom}$ were much smaller than those of $metaFARVAT^{Het}$ in the heterogeneous scenario. In addition, the difference in power between $metaFARVAT^{Hom}$ and $metaFARVAT^{Het}$ increased as the proportion of protective causal variants increased. The most efficient method depends on the disease model, which is often unknown. For example, when all rare causal variants had deleterious effects on the phenotype, burden and VT type *metaFARVAT* outperformed all other approaches, but if there were variants with deleterious and protective effects, SKAT-type *metaFARVAT* was the most efficient. SKAT-O *metaFARVAT* was not always the most powerful, but its empirical power estimates were usually very close to those of the most efficient approach.

The proposed methods can be applied to quantitative phenotypes, and results for quantitative phenotypes are provided in Tables 4.4-4.6 and Figures 4.5–4.7. For quantitative phenotypes, I compared our method with RAREMETAL and seqMeta, since these two methods can only be applied to

quantitative phenotypes. RAREMETAL does not provide the SKAT-O type statistic and seqMeta does not provide the VT type statistic. seqMeta performed better than RAREMETAL in most scenarios and was similar to *metaFARVAT*^{Hom} under homogeneous scenarios. The SKAT-O type statistic in seqMeta did not perform well when there were as many protective variants as deleterious variants in the gene. *metaFARVAT*^{Het} outperformed other methods when the effects of each rare variant differed among studies and when there were variants with deleterious and protective effects within a gene.

Table 4.4 Type I error estimates from simulation study for quantitative phenotypes. The empirical type I error was estimated for proposed methods with 20,000 replicates at the 0.1, 0.01, 10^{-3} and 10^{-4} significance levels for quantitative phenotypes. I applied the proposed methods to meta-analyses based on 3, 6 and 9 studies. I assumed that the number of rare variants is 60, and their MAF <0.1.

Method	*K*	α	Quantitative phenotype			
			Burden	**SKAT**	**SKAT-O**	**VT**
homogeneous *metaFARVAT*	3	0.1	0.0989	0.0965	0.0993	0.1044
		0.01	0.0094	0.0083	0.0094	0.0108
		10^{-3}	0.0012	0.0011	0.0009	0.0018
		10^{-4}	0.0001	0.0001	0.0002	0.0006
	6	0.1	0.1010	0.0948	0.0959	0.0929
		0.01	0.0092	0.0097	0.0094	0.0087
		10^{-3}	0.0010	0.0008	0.0010	0.0009
		10^{-4}	0.0002	0.0001	0.0002	0.0001
	9	0.1	0.0999	0.0963	0.0948	0.0959
		0.01	0.0094	0.0085	0.0090	0.0090
		10^{-3}	0.0008	0.0008	0.0007	0.0006
		10^{-4}	0.0002	0.0000	0.0000	0.0000
heterogeneous *metaFARVAT*	3	0.1	0.1017	0.0984	0.0976	--
		0.01	0.0093	0.0084	0.0085	--
		10^{-3}	0.0010	0.0007	0.0007	--
		10^{-4}	0.0002	0.0001	0.0003	--
	6	0.1	0.1010	0.0996	0.0989	--
		0.01	0.0098	0.0091	0.0095	--
		10^{-3}	0.0011	0.0008	0.0011	--
		10^{-4}	0.0002	0.0001	0.0002	--
	9	0.1	0.1007	0.1072	0.1013	--
		0.01	0.0097	0.0101	0.0094	--
		10^{-3}	0.0011	0.0008	0.0005	--
		10^{-4}	0.0001	0.0001	0.0002	--

The definition of the acronyms in Table 4.4: 1) *K*: the number of studies; 2) SKAT: the sequence kernel association test; 3) Burden: the burden test; 4) SKAT-O: the optimal SKAT; 5) VT: the variable threshold test; 6) α: significance level.

Table 4.5 Empirical power estimates for meta-analyses of quantitative phenotype for homogeneous variants among studies. Empirical power estimates of proposed methods for quantitative phenotypes were calculated with homogeneous effects at the 2.5×10^{-6} significant level. Empirical power estimates of burden, SKAT and VT type of RAREMETAL and seqMeta were calculated with the same dataset and were compared with *metaFARVAT* method.

+/-	Method	3 studies				6 studies				9 studies			
		SKAT	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**
60/0	Fisher	0.6300				0.9870				1.0000			
	minP	0.1155				0.1805				0.2120			
	RAREMETAL	0.0340	0.8330	--	0.7535	0.4930	1.0000	--	1.0000	0.9370	1.0000	--	1.0000
	seqMeta	0.0455	0.8655	0.8715	--	0.5880	1.0000	1.0000	--	0.9650	1.0000	1.0000	--
	Hom	0.0400	0.8650	0.8370	0.8700	0.5720	1.0000	1.0000	1.0000	0.9620	1.0000	1.0000	1.0000
	Het	0.0030	0.8580	0.8605	--	0.0465	1.0000	1.0000	--	0.1850	1.0000	1.0000	--
48/12	Fisher	0.0820				0.4025				0.7240			
	minP	0.0090				0.0185				0.0185			
	RAREMETAL	0.0295	0.1065	--	0.0830	0.4750	0.5935	--	0.5215	0.9285	0.9120	--	0.8770
	seqMeta	0.0440	0.1475	0.1640	--	0.5740	0.6555	0.6855	--	0.9545	0.9360	0.9525	--
	Hom	0.0400	0.1455	0.2345	0.1875	0.5580	0.6540	0.8705	0.6855	0.9505	0.9335	0.9955	0.9395
	Het	0.0070	0.1465	0.2315	--	0.0750	0.6440	0.7810	--	0.2070	0.9295	0.9725	--
30/30	Fisher	0.0035				0.0225				0.0690			
	minP	0.0000				0.0000				0.0005			
	RAREMETAL	0.0420	0.0000	--	0.0010	0.4470	0.0000	--	0.0025	0.9140	0.0000	--	0.0110
	seqMeta	0.0590	0.0000	0.0000	--	0.5610	0.0000	0.0000	--	0.9500	0.0000	0.0000	--

	Hom	0.0520	0.0000	0.0250	0.0025	0.5515	0.0000	0.4065	0.0070	0.9470	0.0000	0.8975	0.0250
	Het	0.0125	0.0000	0.0050	--	0.0845	0.0000	0.0530	--	0.2430	0.0000	0.1535	--
	Fisher	0.1725				0.6295				0.9005			
	minP	0.0195				0.0270				0.0335			
	RAREMETAL	0.0540	0.2405	--	0.1925	0.6015	0.8490	--	0.7990	0.9635	0.9885	--	0.9795
	seqMeta	0.0605	0.2820	0.3095	--	0.6510	0.8840	0.9015	--	0.9720	0.9910	0.9945	--
	Hom	0.0535	0.2865	0.3900	0.3410	0.6385	0.8820	0.9490	0.8910	0.9705	0.9915	1.0000	0.9915
	Het	0.0100	0.2830	0.3905	--	0.0935	0.8735	0.9380	--	0.2750	0.9900	0.9970	--
24/6	Fisher	0.0175				0.1295				0.3445			
	minP	0.0025				0.0025				0.0025			
	RAREMETAL	0.0515	0.0100	--	0.0070	0.5460	0.1145	--	0.0995	0.9465	0.3650	--	0.3290
	seqMeta	0.0620	0.0150	0.0200	--	0.6155	0.1490	0.1915	--	0.9595	0.4335	0.5120	--
	Hom	0.0565	0.0145	0.0825	0.0315	0.6060	0.1490	0.6478	0.2045	0.9565	0.4305	0.9720	0.4805
	Het	0.0130	0.0125	0.0615	--	0.0795	0.1410	0.3690	--	0.2335	0.4120	0.7245	--
15/15	Fisher	0.0010				0.0185				0.0655			
	minP	0.0000				0.0005				0.0000			
	RAREMETAL	0.0360	0.0000	--	0.0000	0.4420	0.0000	--	0.0010	0.9065	0.0000	--	0.0070
	seqMeta	0.0485	0.0000	0.0000	--	0.5525	0.0000	0.0000	--	0.9470	0.0000	0.0000	--
	Hom	0.0470	0.0000	0.0190	0.0000	0.5420	0.0000	0.3875	0.0070	0.9470	0.0000	0.8825	0.0270
	Het	0.0070	0.0000	0.0035	--	0.0810	0.0000	0.0515	--	0.2420	0.0000	0.1570	--

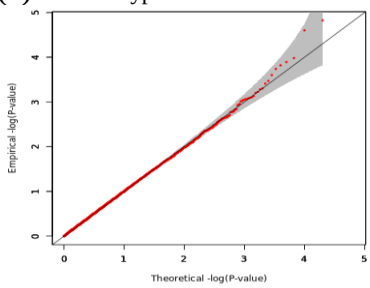
Table 4.6 Empirical power estimates for meta-analyses of quantitative phenotype for heterogeneous variants among studies. Empirical power estimates of burden, SKAT, SKAT-O and VT type of *metaFARVAT*^{Hom} and *metaFARVAT*^{Het} were calculated for quantitative phenotypes with heterogeneous effects at the 2.5×10^{-6} significant level. Empirical power estimates of burden, SKAT and VT type of RAREMETAL and seqMeta were calculated with the same dataset and were compared with *metaFARVAT* method.

+/-	Method	3 studies				6 studies				9 studies			
		SKAT	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**
48/12	Fisher	0.0720				0.3765				0.7120			
	minP	0.0080				0.0085				0.0130			
	RAREMETAL	0.0190	0.2675	--	0.2015	0.1405	0.8720	--	0.8050	0.4410	0.9910	--	0.9825
	seqMeta	0.0095	0.1270	0.1410	--	0.0560	0.6205	0.6400	--	0.2345	0.9345	0.9390	--
	Hom	0.0085	0.1270	0.1515	0.1740	0.0580	0.6190	0.6525	0.6675	0.2290	0.9335	0.9410	0.9465
	Het	0.0045	0.1205	0.2040	--	0.0615	0.6065	0.7650	--	0.2075	0.9255	0.9700	--
30/30	Fisher	0.0050				0.0190				0.0565			
	minP	0.0000				0.0000				0.0005			
	RAREMETAL	0.0015	0.0000	--	0.0000	0.0030	0.0000	--	0.0005	0.0015	0.0000	--	0.0000
	seqMeta	0.0020	0.0005	0.0005	--	0.0015	0.0000	0.0000	--	0.0020	0.0000	0.0000	--
	Hom	0.0020	0.0005	0.0005	0.0015	0.0015	0.0000	0.0005	0.0005	0.0025	0.0000	0.0005	0.0000
	Het	0.0130	0.0005	0.0070	--	0.0750	0.0000	0.0440	--	0.2265	0.0000	0.1370	--
30/0	Fisher	0.1825				0.6635				0.9060			
	minP	0.0230				0.0295				0.0350			
	RAREMETAL	0.0190	0.2675	--	0.2015	0.1405	0.8720	--	0.8050	0.4410	0.9910	--	0.9825
	seqMeta	0.0135	0.3025	0.3175	--	0.1680	0.8875	0.8955	--	0.4975	0.9945	0.9965	--

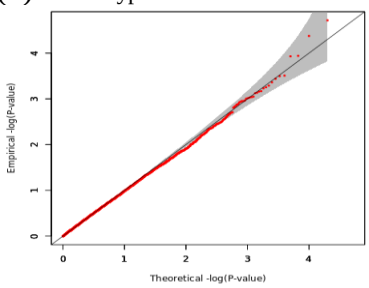
	Hom	0.0125	0.3005	0.3245	0.3540	0.1575	0.8855	0.8875	0.8995	0.4800	0.9950	0.9925	0.9945
	Het	0.0115	0.2865	0.4045	--	0.0960	0.8770	0.9295	--	0.2530	0.9915	0.9985	--
24/6	Fisher	0.0215				0.1395				0.3550			
	minP	0.0025				0.0035				0.0050			
	RAREMETAL	0.0035	0.0180	--	0.0110	0.0225	0.1595	--	0.1080	0.0575	0.4230	--	0.3390
	seqMeta	0.0035	0.0225	0.0245	--	0.0190	0.1745	0.1900	--	0.0600	0.4435	0.4630	--
	Hom	0.0030	0.0220	0.0360	0.0395	0.0175	0.1710	0.2270	0.2240	0.0600	0.4430	0.5000	0.5090
	Het	0.0075	0.0200	0.0605	--	0.0750	0.1560	0.3715	--	0.2460	0.4110	0.7275	--
15/15	Fisher	0.0020				0.0200				0.0615			
	minP	0.0005				0.0000				0.0000			
	RAREMETAL	0.0025	0.0000	--	0.0000	0.0015	0.0000	--	0.0000	0.0035	0.0000	--	0.0000
	seqMeta	0.0040	0.0000	0.0000	--	0.0005	0.0000	0.0000	--	0.0035	0.0000	0.0000	--
	Hom	0.0035	0.0000	0.0000	0.0010	0.0005	0.0000	0.0005	0.0000	0.0040	0.0000	0.0005	0.0015
	Het	0.0130	0.0000	0.0090	--	0.0700	0.0000	0.0485	--	0.2380	0.0000	0.1485	--

The definition of the acronyms in Table 4.6: 1) +/-: the number of causal variants with positive and negative effect; 2) SKAT: the sequence kernel association test; 3) Burden: the burden test; 4) SKAT-O: the optimal SKAT; 5) VT: the variable threshold test; 6) Fisher: Fisher's method; 7) minP: the minimum p-value method; 8) Hom: homogeneous *metaFARVAT*; 9) Het: heterogeneous *metaFARVAT*.

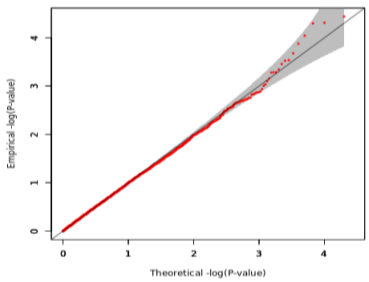
(a) Burden-type *metaFARVAT*^{Hom}



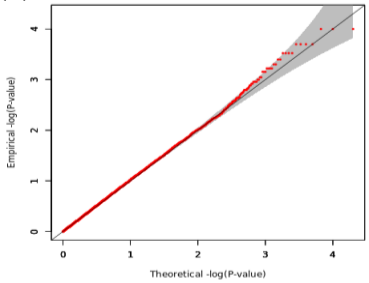
(b) SKAT-type *metaFARVAT*^{Hom}



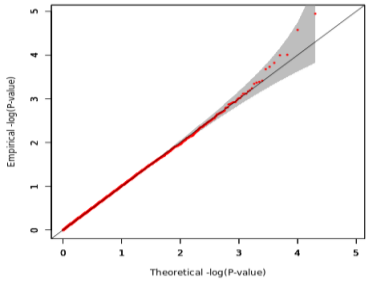
(c) SKATO-type *metaFARVAT*^{Hom}



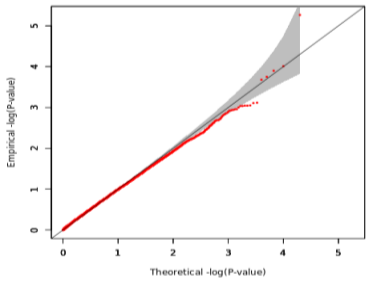
(d) VT-type *metaFARVAT*^{Hom}



(e) Burden-type *metaFARVAT*^{Het}



(f) SKAT-type *metaFARVAT*^{Het}



(g) SKATO-type *metaFARVAT*^{Het}

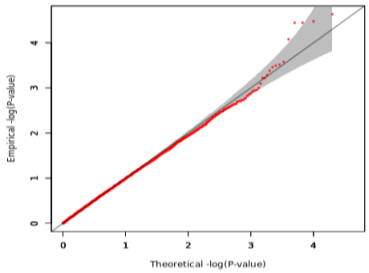


Figure 4.5 QQ plots for meta-analyses of quantitative phenotype based on 3 studies. QQ plots were provided for results from the proposed methods under the null hypothesis. The empirical p-values were calculated under the null hypothesis with 20,000 replicates.

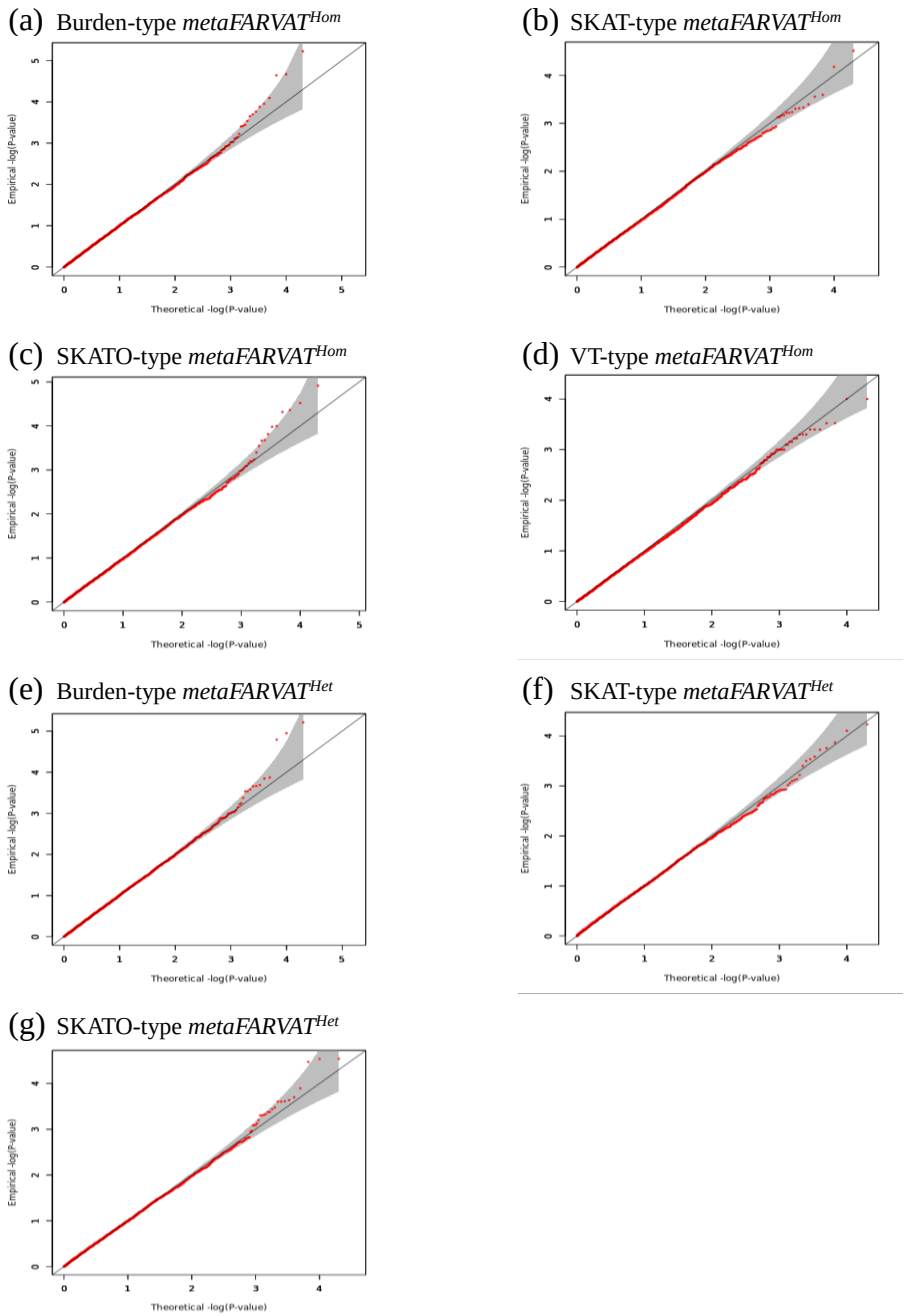
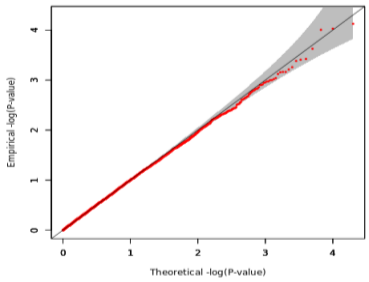
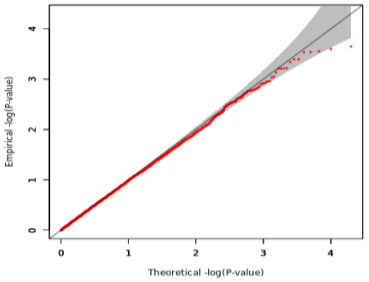


Figure 4.6 QQ plots for meta-analyses of quantitative phenotype based on 6 studies. QQ plots were provided for results from the proposed methods under the null hypothesis. The empirical p-values were calculated under the null hypothesis with 20,000 replicates.

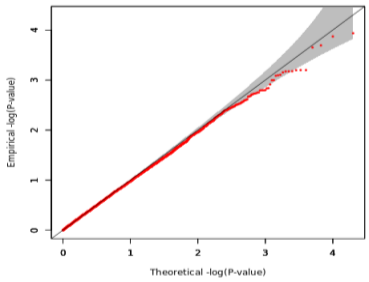
(a) Burden-type *metaFARVAT*^{Hom}



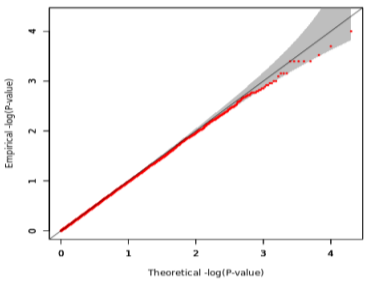
(b) SKAT-type *metaFARVAT*^{Hom}



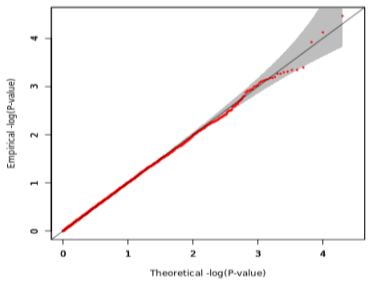
(c) SKATO-type *metaFARVAT*^{Hom}



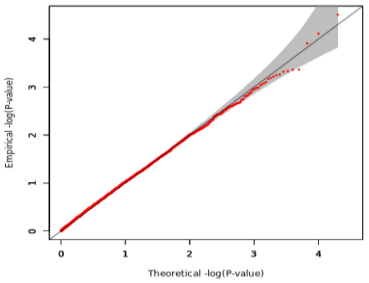
(d) VT-type *metaFARVAT*^{Hom}



(e) Burden-type *metaFARVAT*^{Het}



(f) SKAT-type *metaFARVAT*^{Het}



(g) SKATO-type *metaFARVAT*^{Het}

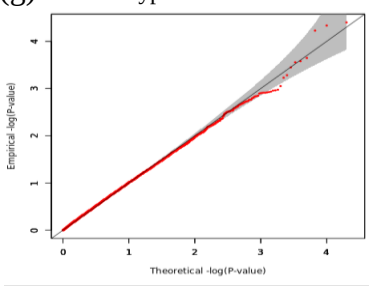


Figure 4.7 QQ plots for meta-analyses of quantitative phenotype based on 9 studies. QQ plots were provided for results from the proposed methods under the null hypothesis. The empirical p-values were calculated under the null hypothesis with 20,000 replicates.

4.4 Application to COPD data

I considered previously reported family-based WES data from Boston Early-Onset COPD Study (EOCOPD) and COPDGene case-control subjects for meta-analysis (Qiao et al. 2016). Details of the EOCOPD study have been described previously (Silverman et al. 1998). The EOCOPD data are derived from an extended pedigree-based design. Probands were 53 years old or younger with prebronchodilator forced expiratory volume in one second (FEV_1) of ≤40%, physician-diagnosed COPD, and without severe alpha-1 antitrypsin deficiency. All first-degree relatives, older second-degree relatives, and additional affected family members were enrolled. There were 49 pedigrees with at least 2 affected family members selected for WES. COPDGene was a multi-center study of smokers with and without COPD and included African-Americans and non-Hispanic whites (Regan et al. 2010). The COPDGene participants, consisting of 10,192 smokers, had at least 10 pack years of smoking, and their ages were between 45 and 80 years. From the COPDGene study, 204 COPD subjects with GOLD spirometry grades 3–4 (post-bronchodilator FEV_1<50% and ratio of FEV_1 to forced vital capacity (FEV_1/FVC) <0.7), as well as 195 controls with normal spirometry (frequency-matched to COPD cases on pack-years of cigarette smoking), were chosen for WES.

Sequencing for both cohorts was performed at the University of Washington (Seattle, WA), using Nimblegen V2 capture (Roche NimbleGen, Inc., Madison, WI) and the Illumina platform (Illumina, Inc., San Diego, CA).

Participants selected from the COPDGene cohort were sequenced via the NHLBI Exome Sequencing Program, and EOCOPD subjects were sequenced as part of the Center for Mendelian Genomics. Quality control (QC) filtering for both data sets was performed by the method of Qiao et al (Qiao et al. 2016) and filtered out variants with Mendelian errors (for family-based data), call rate <99%, HWE p-value $<10^{-8}$, and average sequencing depth <12, as well as excluding subjects with pedigree, racial, or sex mismatches. After QC, there were 303 individuals from 49 families and 124,288 variants in the EOCOPD data set, and there were 394 unrelated individuals and 108,443 variants in the COPDGene data set. For rare variant analyses, I assumed that variants with MAFs <5% in dbSNP were rare, and in both studies, I separately filtered out singleton variants or genes with MACs less than 10. Finally, 88,737 rare variants in 13,935 genes were analyzed in the EOCOPD data set, and 24,846 rare variants in 10,550 genes were tested in the COPDGene data set. For both EOCOPD and COPDGene data, GRMs were estimated for variants with MAFs >5% and were incorporated as variance-covariance matrices of genotypes to adjust for population substructure. Effects of covariates for binary phenotypes were adjusted by using the BLUP as an offset. First, I fitted the linear mixed model with adjustments for age, sex, and pack-years of smoking as covariates, and then BLUP was set as the offset for the proposed methods. A description of the two datasets is provided in Table 4.7.

Table 4.7 The description of chronic obstructive pulmonary disease (COPD) datasets, EOCOPD WES and COPDGene. This description includes the range of age, and the number of individuals, families, females/males, cases/controls/missing, variants, rare variants (MAF <5% in dbSNP) and genes.

	EOCOPD WES	**COPDGene**
Age	[21, 87]	[46, 81]
Sample size	303	394
Families	49	--
F/M	209/138	211/200
Cases/controls	155/148	204/195
Variants	124,288	108,443
Rare variants	88,373	24,846
Genes	13,935	10,550

The definition of the acronyms in Table 4.7: 1) Families: the number of families; 2) F/M: the number of females and males; 3) Cases/controls: the number of cases and controls; 4) Variants: the number of variants; 5) Rare variants: the number of rare variants; 6) Genes: the number of genes.

To identify rare variants associated with COPD, I separately conducted rare variant analyses with EOCOPD and COPDGene data. Manhattan and QQ plots are provided in Figure 4.8. According to the results, there were no exome-wide significant genes. I also conducted meta-analysis with metaFARVAT^{Hom} and metaFARVAT^{Het}. For both statistics, v_1 and v_2 were set to 1. The QQ plots in Figure 4.9 show that SKAT-O type metaFARVAT^{Het} and metaFARVAT^{Hom} preserved the nominal significance level. However, VT type metaFARVAT exhibited some inflation, and its results are therefore not included in Table 4.8. Manhattan plots are provided in Figure 4.10. The Bonferroni-corrected 0.05 genome-wide significance level was 6.76×10^{-6} and is indicated by a solid blue line. Table 4 shows that *DLEC1* achieved genome-wide significance under both methods, and *ZNF441* was implicated with potentially significant results (p-value $<10^{-4}$) by metaFARVAT^{Hom} SKAT-O. *DLEC1* is a protein-coding gene encoding a cilia and flagella-associated protein. Downregulation of this gene has been observed in several human cancers, including lung, esophageal, and renal tumors and head and neck squamous cell carcinoma. It has also been found that reduced expression of this gene in tumor cells is a result of aberrant promoter methylation. Several alternatively spliced transcripts have been observed that contain disrupted coding regions and likely encode nonfunctional proteins (Pruitt et al. 2016). The clinical conditions include alveolar cell carcinoma, chest pain, and lymphadenopathy.

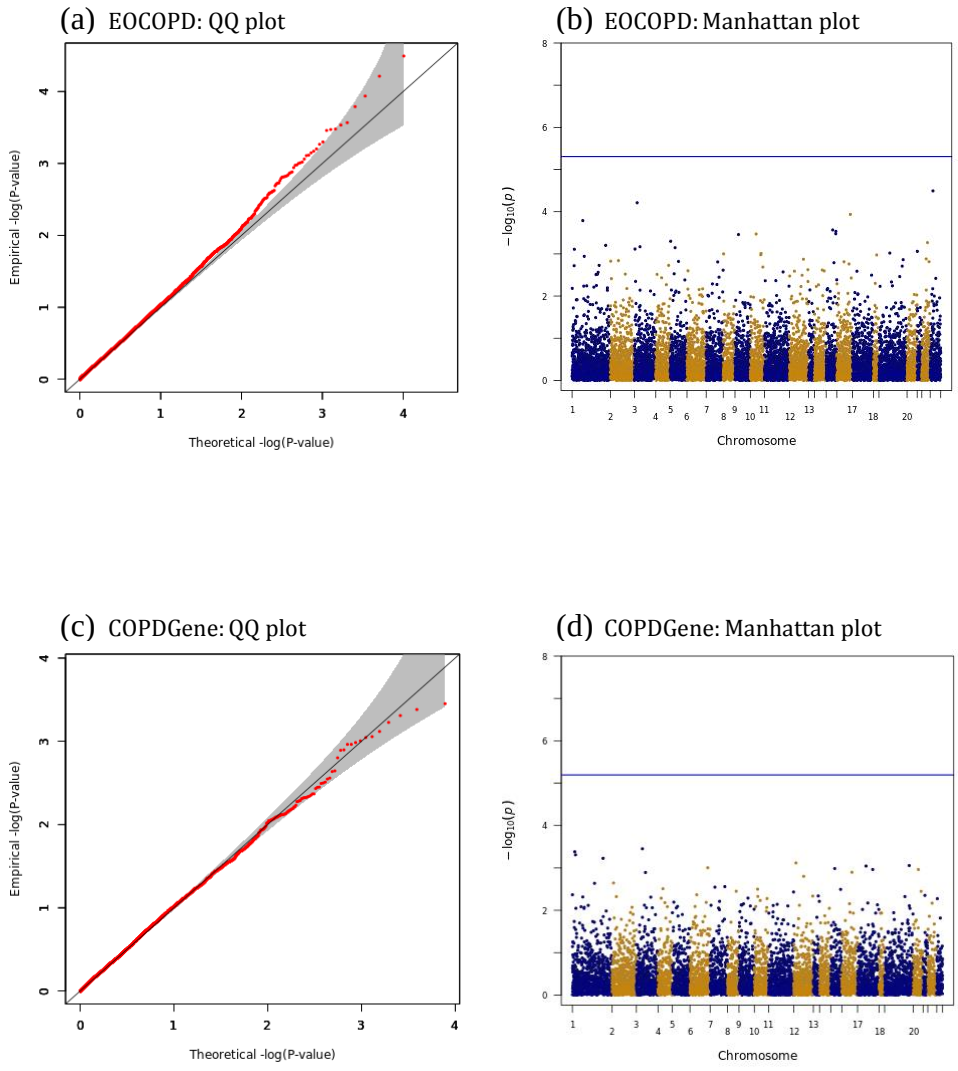


Figure 4.8 QQ plots and Manhattan plots are based on the results of the association analyses with EOCOPD and COPDGene datasets using *FARVAT*. EOCOPD and COPDGene were separately analyzed using FARVAT and the results of SKAT-O type statistic were used for QQ plots and Manhattan plots. (a) and (b) are for EOCOPD dataset, and (c) and (d) are for COPDGene.

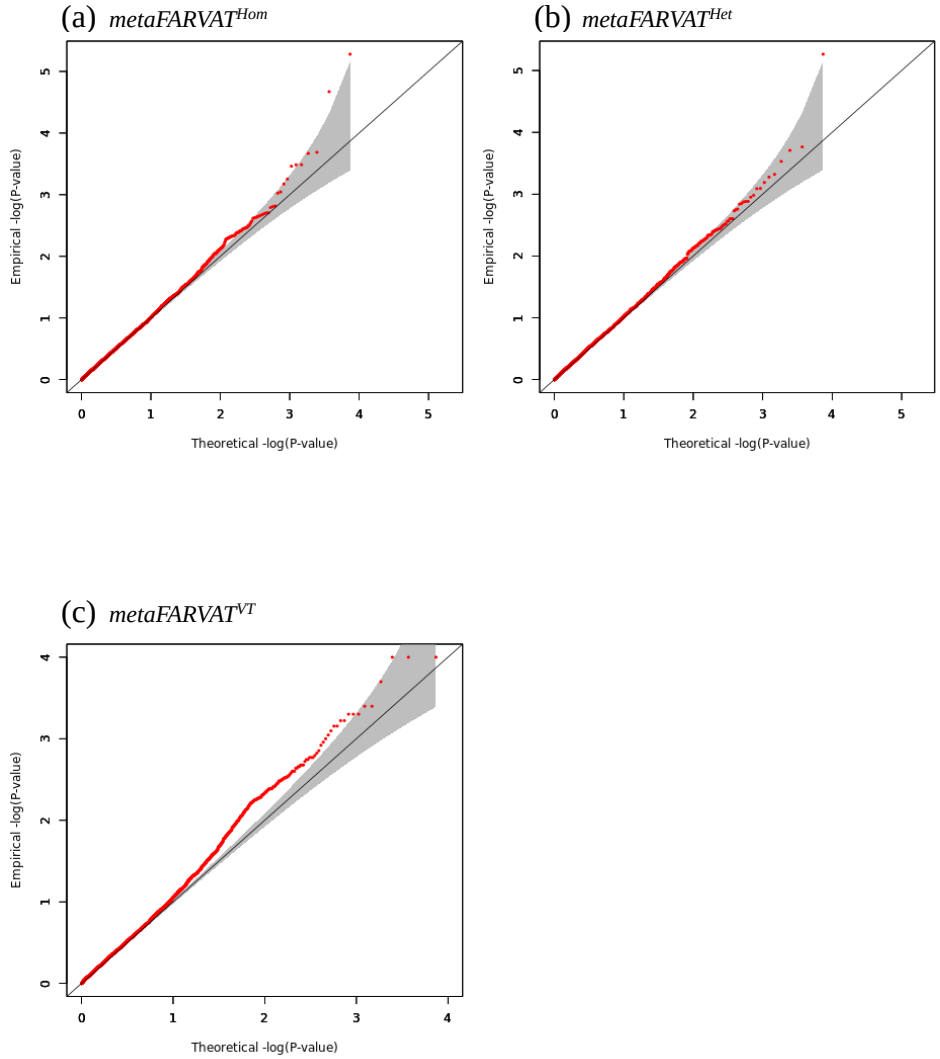


Figure 4.9 QQ plots of results from *metaFARVAT* with the EOCOPD and the COPDGene datasets. *metaFARVAT* was applied to meta-analysis with the EOCOPD and the COPDGene datasets. (a), (b) and (c) were based on results from SKAT-O *metaFARVAT*^{Hom}, SKAT-O *metaFARVAT*^{Het} and *metaFARVAT*^{VT} respectively.

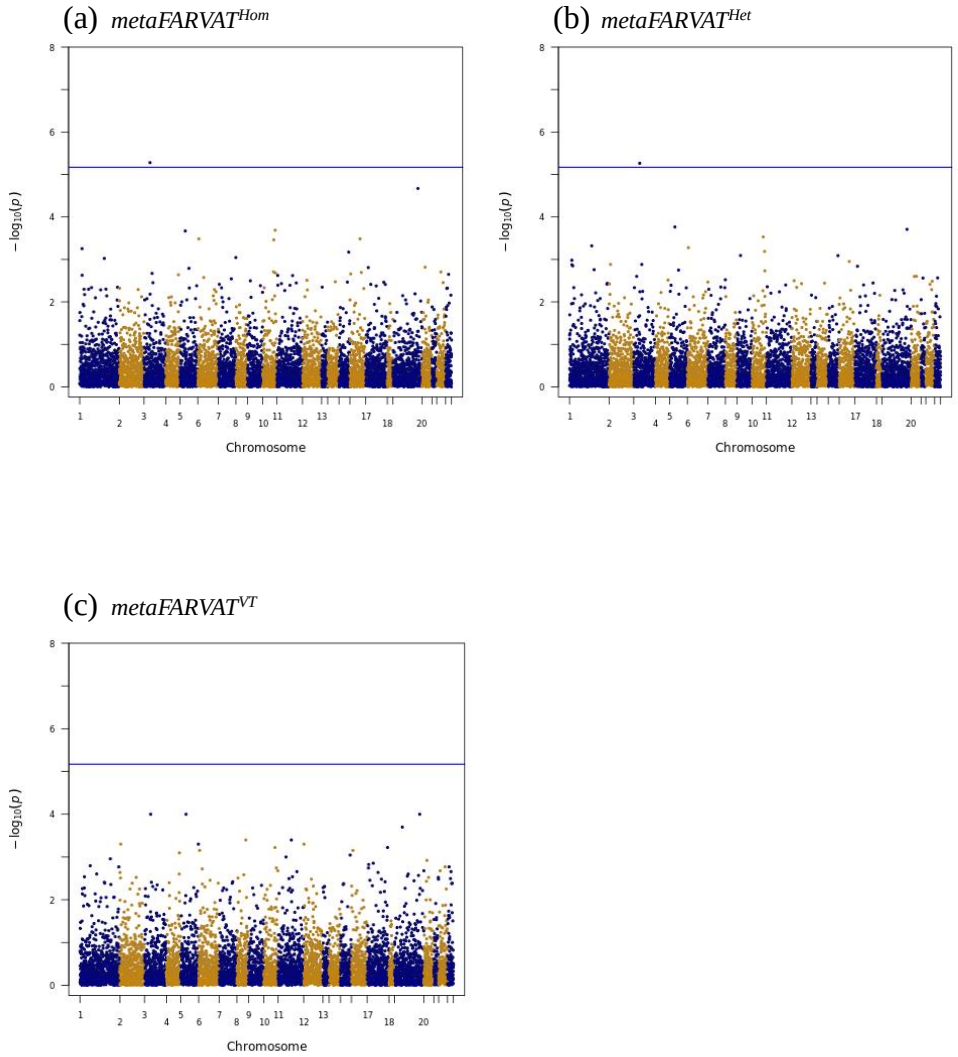


Figure 4.10 Manhattan plots of results from *metaFARVAT* with the EOCOPD and the COPDGene datasets. *metaFARVAT* was applied to meta-analysis with the EOCOPD and the COPDGene datasets. (a), (b) and (c) were based on results from SKAT-O *metaFARVAT*^{Hom}, SKAT-O *metaFARVAT*^{Het} and *metaFARVAT*^{VT} respectively.

Table 4.8 The candidate genes found by meta-analysis in chronic obstructive pulmonary disease (COPD) studies.

Method	Data	Gene	Sample size	Chr	Start	End	#rare variant	MAC	P_B	P_S	*c*	P_O
metaFARVAT ^{Hom}	EOCOPD & COPDGene	*DLEC1*	697	3	38080978	38163785	9	66	1.21e-05	1.02e-04	0.25	5.24e-06
		ZNF441	697	19	11890983	11892255	2	24	9.88e-05	3.46e-04	1	2.13e-05
metaFARVAT ^{Het}	EOCOPD & COPDGene	*DLEC1*	697	3	38080978	38163785	15	66	8.03e-06	9.54e-04	0.16	5.43e-06
FARVAT	EOCOPD	*DLEC1*	303	3	38080978	38163785	9	28	3.70e-03	1.47e-02	1	7.24e-03
		ZNF441	303	19	11890983	11892255	2	13	1.12e-03	3.83e-03	1	1.37e-03
FARVAT	COPDGene	*DLEC1*	394	3	38080978	38163785	6	38	5.80e-04	7.96e-04	0.25	3.53e-04
		ZNF441	394	19	11890983	11892255	1	11	3.09e-02	3.09e-02	1	3.09e-02

The definition of the acronyms in Table 4.8: 1) Chr: chromosome; 2) #rare variants: the number of rare variants in the gene; 3) MAC: minor allele count; 4) P_B: the p-value of burden type test; 5) P_S: the p-value of SKAT type test; 6) *c*: the parameter used for SKAT-O; 7) P_O: the p-value of SKAT-O type test; 8) Start/End: start position and end position.

To access the performance of *metaFARVAT* for case-control designs, I applied it to two case-control dataset from COPDGene study and compared it with metaSKAT method (Lee et al. 2013). The COPDGene study is a multi-center epidemiologic and genetic study of 10,192 current or ex-smokers. COPDGene subjects were sequenced in two sets. The first set sequenced at Baylor included severe COPD cases, The Global Initiative for Chronic Obstructive Lung Disease (GOLD) Grades 3 or 4 (post-bronchodilator FEV1 < 50% predicted and FEV1/FVC < 0.70), with no age requirement. Controls were selected to be resistant smokers with normal lung function with ages >55 years. The second set sequenced as part of the NHLBI Exome Sequencing Project (ESP) included severe COPD cases with GOLD Grades 3 or 4, and aged < 65 years old, with substantial emphysema (>15% at −950 HU) by quantitative chest CT scan. Controls were selected to be resistant smokers with frequency-matched pack-years of cigarette smoking, normal lung function (FEV_1 > 80% predicted and FEV_1/FVC > 70%), aged > 65 years old and no significant emphysema (< 5% at −950 HU). All subjects were sequenced using Nimblegen capture and Illumina platforms. The COPDGene ESP subjects were all sequenced at the University of Washington, using Nimblegen V2 exome capture; COPDGene Baylor samples used VChrome capture. Alignment, variant calling and quality control were performed using bwa, GATK and in-house pipelines, respectively. As COPDGene ESP and COPDGene Baylor used slightly different capture platforms, calling was performed on these datasets separately (Qiao et al. 2018). The description of two datasets are summarized

in Table 4.9. The sample sizes are 609 and 380 in each data. There are 293 cases and 316 controls in Baylor data, and 192 cases and 188 controls in ESP data. I used COPD status as binary phenotype and ratio (FEV1/FEV) as quantitative phenotype, and adjusted both phenotypes using covariates: gender, packs per year, height and age. Variants with moderate and high impact in SnpEff and MAF<5% in 1000 Genome were selected. After filtering, there are 54,724 variants in 13,982 genes for Baylor data, and 46,709 in 13,335 genes for ESP data.

Table 4.9 The description of chronic obstructive pulmonary disease (COPD) datasets, Baylor and ESP. This description includes the range of age, and the number of samples, females/males, cases/controls, variants and genes.

Description		Baylor	ESP
Sample size		609	380
Cases/controls		293/316	192/188
Phenotypes	fev1 %	80.60 (61.80)	49.40 (67.98)
	ratio %	70.00 (41.00)	55.00 (45.00)
Covariates	Gender: M/F	346/263	185/195
	Packs/year	50.90 (24.40)	45.00 (23.03)
	Height(cm)	169.20 (14.00)	168.80 (13.33)
	Age	64.70 (7.80)	62.80 (11.23)
Gene sets	Variants	54,727	46,709
	Genes	13,982	13,335

The definition of the acronyms in Table 4.9: 1) FEV1: bronchodilator forced expiratory volume in one second; 2) ratio: FEV1/FVC; FEV1 over forced vital capacity. 3) Packs/year: the number of cigarette packs that the subjects smoke per year. 4) Median(IQR): median and interquartile range (IQR) of the quantitative variables. 5) Variants: the variants with moderate or high impact in SnpEff and minor allele frequency <5% in 1000 Genome were selected.

The meta-analyses for Baylor and ESP data were conducted using *metaFARVAT* and metaSKAT SKAT-O method. The QQ-plots and Manhattan plots for *metaFARVAT* and metaSKAT with quantitative phenotype, ratio, are displayed in Figure 4.11 and 4.12, respectively, those with binary phenotype, COPD status are shown in Figure 4.13 and 4.14. It shows that both methods are statistically valid for dichotomous and quantitative phenotypes. There is one significant result, *PLD5*, detected by metsSKAT with affected status. *PLD5* is a protein coding gene, which encodes protein Inactive phosphatidylcholine-hydrolyzing phospholipase D5 and is highly expressed in pigmented layer of retina. However, there is no evidence showing *PLD5* is strongly expressed in lungs. To compare the two methods, I selected the genes proven related to ratio and COPD respectively from GWAS catalog and provided p-value plots in Figure 4.15, of which x axis is $-\log p_1$, where p_1 is the p-value of *metaFARVAT* SKAT-O statistic, and y axis is $-\log p_2$, where p_2 is the p-value of *metaSKAT* SKAT-O statistic. It shows that the coefficients of x are 0.897 and 0.877 for homogeneous and heterogeneous model with ratio, and those are 0.559 and 0.647 with COPD. The p-values of these coefficients are extremely small. Therefore, I can conclude that, for quantitative phenotype, the p-values from the two methods are similar but those from *metaFARVAT* are slightly smaller, and for binary phenotype, the difference between the p-values from the two methods are larger and *metaFARVAT* performs better. There are three main reasons for the difference in performance: 1) *metaFARVAT* considers genetic relatedness using generalized linear mixed model with GRM even for case-

control data, which effectively reduce the heterogeneity due to population substructure and admixture. metaSKAT does not consider this and it used generalized linear model. 2) *metaFARVAT* uses quasi-likelihood and adjusts genotype with its best linear unbiased estimator, while metaSKAT uses genotype directly. 3) To adjust phenotype with covariates, *metaFARVAT* uses linear model and metaSKAT uses logistic model.

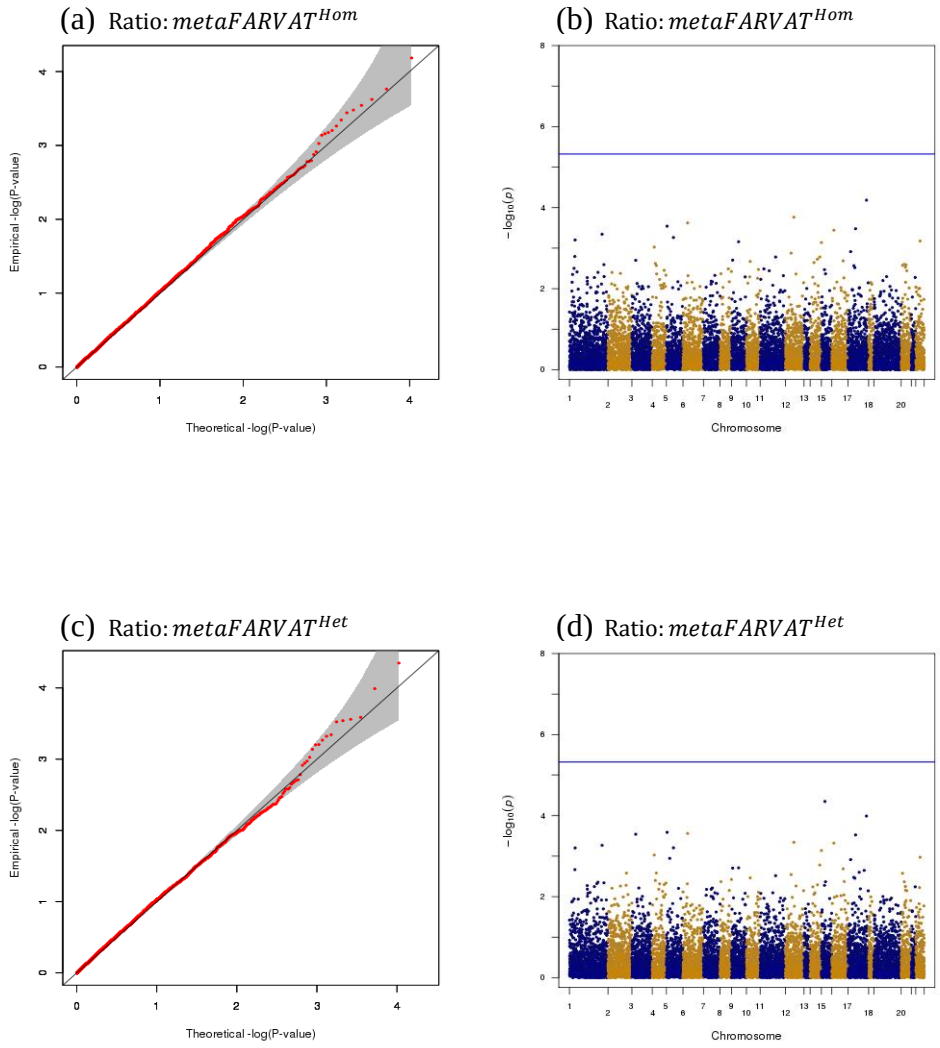


Figure 4.11 QQ plots and Manhattan plots for meta-analysis with ratio using homogeneous and heterogeneous *metaFARVAT*. *metaFARVAT* was applied to meta-analysis of the COPDGene Baylor and ESP datasets with the quantitative phenotype, ratio. (a), (b) were based on the results from SKAT-O *metaFARVAT*^{Hom}, and (c), (d) were based on the results from SKAT-O *metaFARVAT*^{Het}.

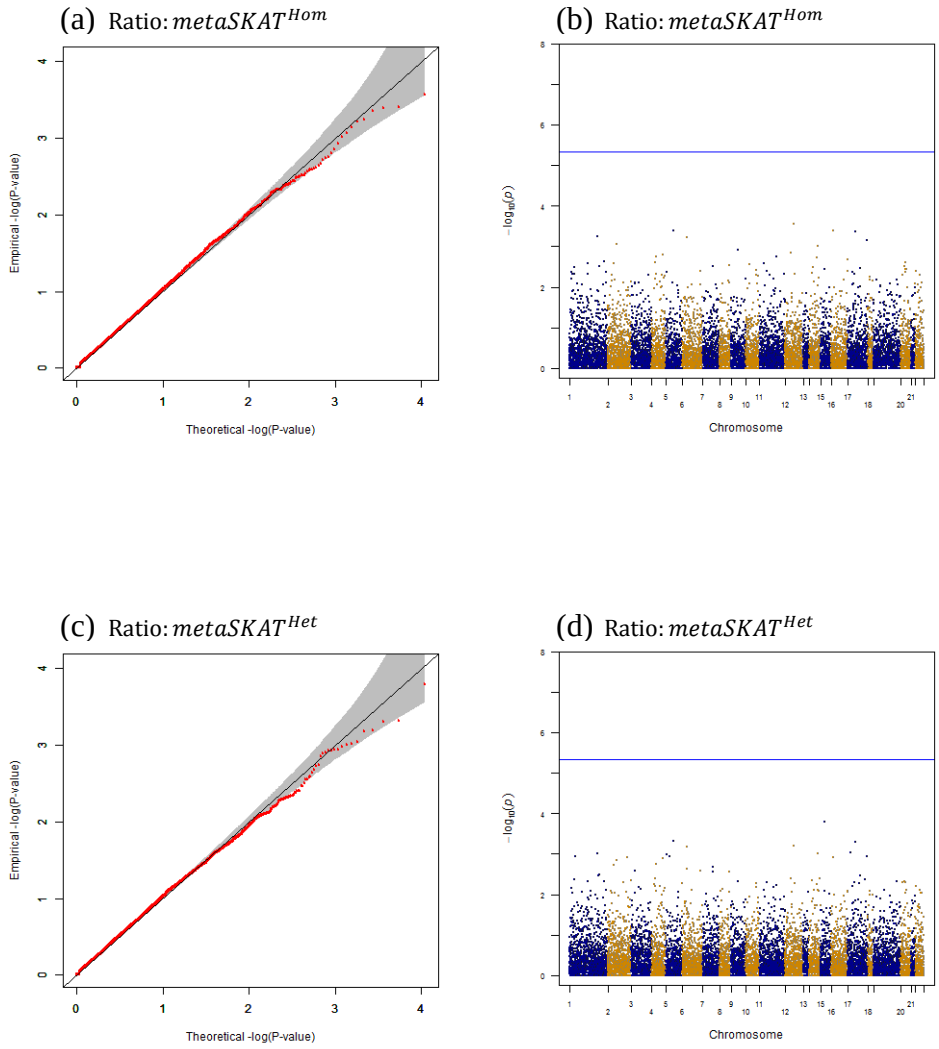


Figure 4.12 QQ plots and Manhattan plots for meta-analysis with ratio using homogeneous and heterogeneous *metaSKAT*. *metaSKAT* was applied to meta-analysis of the COPDGene Baylor and ESP datasets with the quantitative phenotype, ratio. (a), (b) were based on the results from SKAT-O *metaSKAT*^{Hom}, and (c), (d) were based on the results from SKAT-O *metaSKAT*^{Het}.

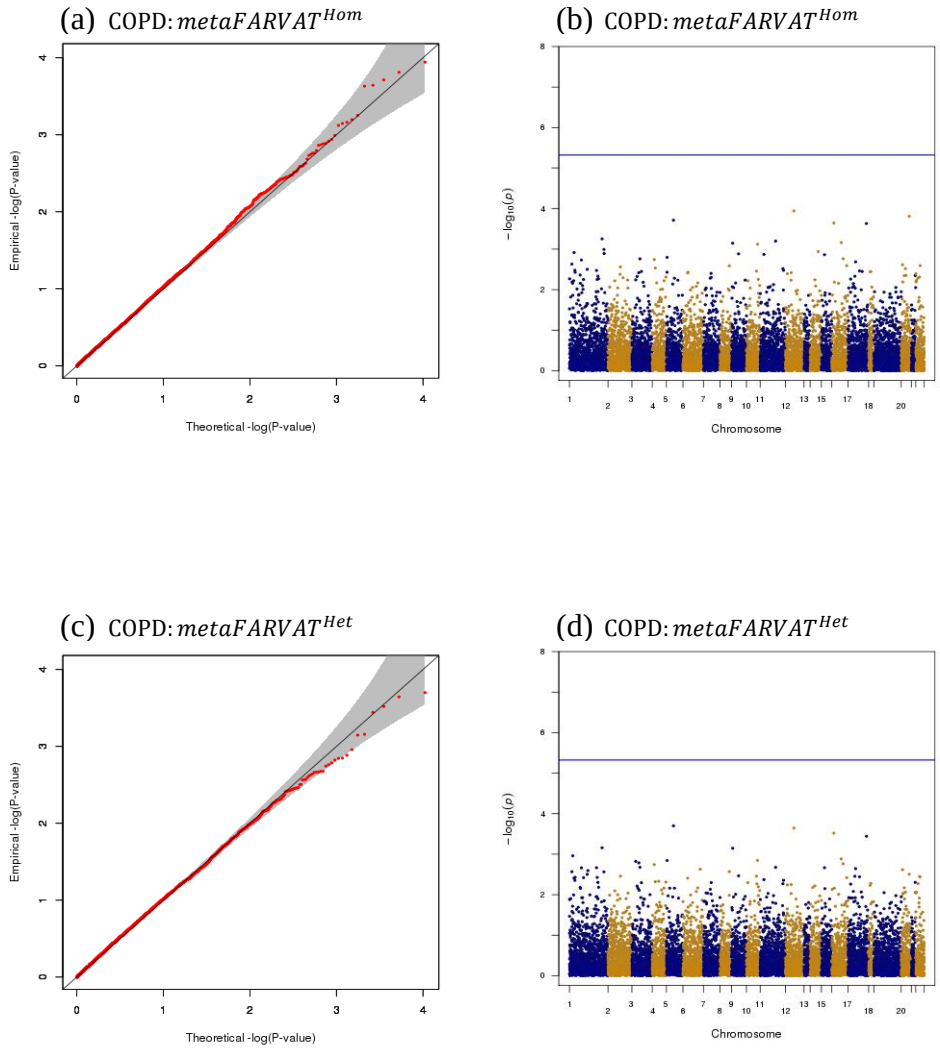


Figure 4.13 QQ plots and Manhattan plots for meta-analysis with COPD status using homogeneous and heterogeneous *metaFARVAT*. *metaFARVAT* was applied to meta-analysis of the COPDGene Baylor and ESP datasets with the dichotomous phenotype, COPD status. (a), (b) were based on the results from SKAT-O *metaFARVAT*^{Hom}, and (c), (d) were based on the results from SKAT-O *metaFARVAT*^{Het}.

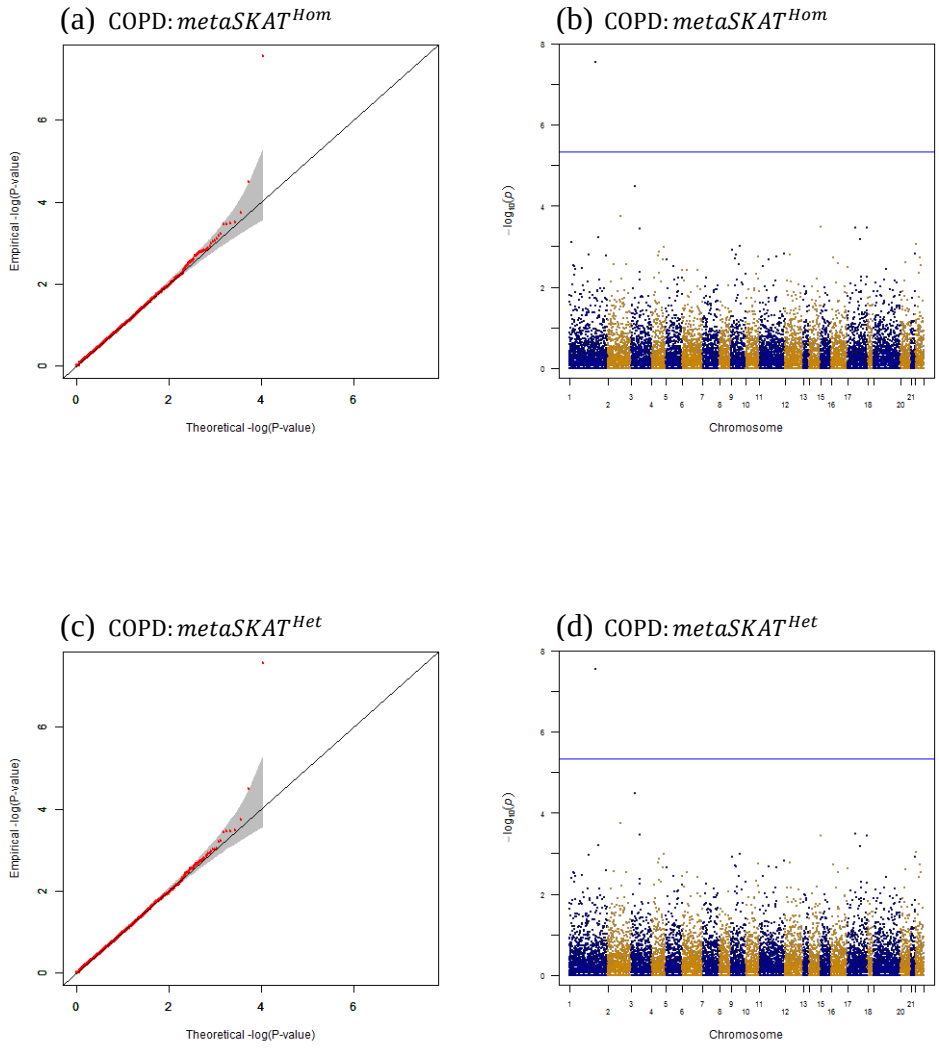


Figure 4.14 QQ plots and Manhattan plots for meta-analysis with COPD status using homogeneous and heterogeneous *metaSKAT*. *metaSKAT* was applied to meta-analysis of the COPDGene Baylor and ESP datasets with the dichotomous phenotype, COPD status. (a), (b) were based on the results from SKAT-O *metaSKAT*^{Hom}, and (c), (d) were based on the results from SKAT-O *metaSKAT*^{Het}.

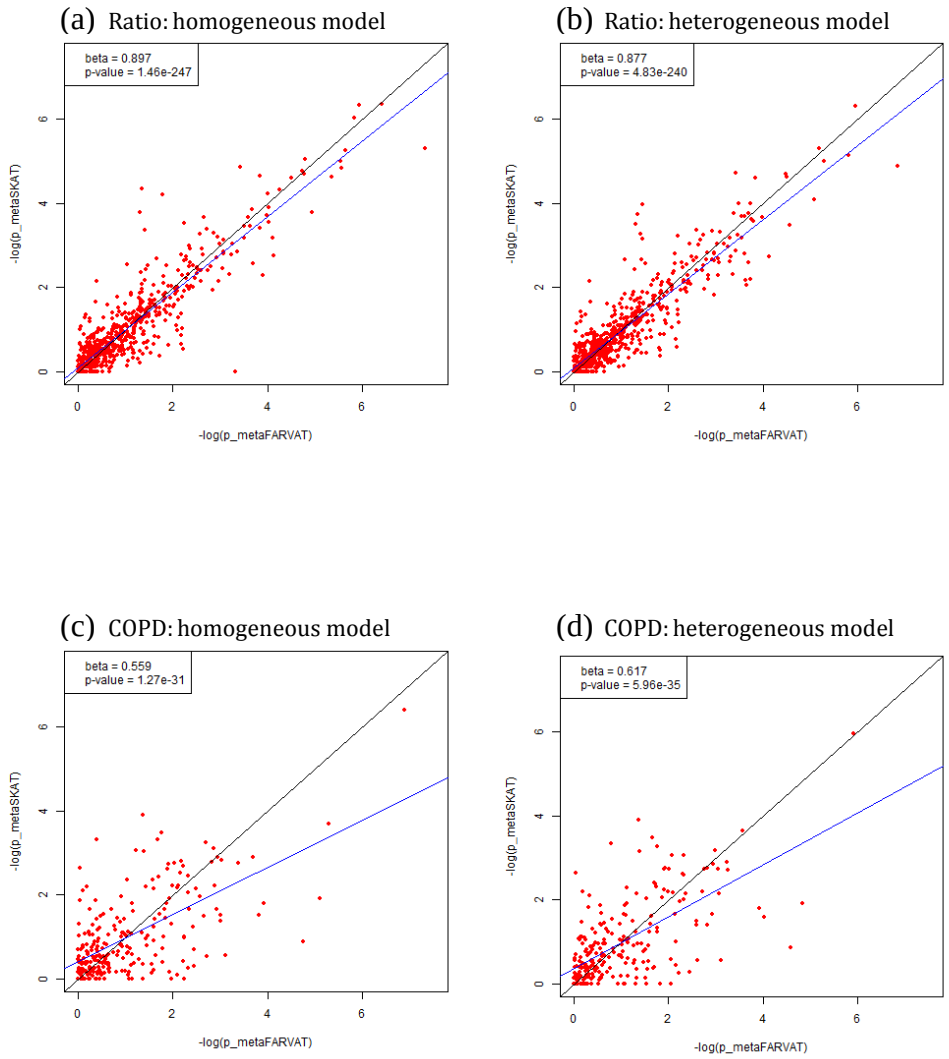


Figure 4.15 *P*-value plots for meta-analysis with *metaFARVAT* and metaSKAT. Homogeneous and heterogeneous models of both methods were applied to meta-analysis of the COPDGene Baylor and ESP datasets with ratio and COPD status. X axis is $-\log p_1$, where p_1 is the p-value of SKAT-O *metaFARVAT*, and y axis is $-\log p_2$, where p_2 is the p-value of SKAT-O metaSKAT. (a), (b) were based on the results from ratio, and (c), (d) were based on the results from COPD status. The blue line is a linear regression line for x and y axis. The coefficients and their p-values are shown in the legends.

4.5 Discussion

Family-based association methods are robust against population substructure, and because of genetic homogeneity among family members, they are often utilized for rare variant association analyses. Multiple approaches have been proposed, and Tang and Lin (Tang and Lin 2015) provided a comprehensive overview of the statistical methods for meta-analysis of sequencing studies for discovering rare variant associations. According to their overview, RAREMETAL (Feng et al. 2014, Liu et al. 2014) and seqMeta (Chen et al. 2014) can be applied to family-based samples. However, these methods can consider only homogeneous effects with quantitative phenotypes, and no statistical methods for dichotomous phenotypes with family-based samples have been proposed.

In this study, I proposed a new meta-analysis method for family-based rare variant association analyses with dichotomous phenotypes, which can test both homogeneous and heterogeneous effects of variants in different studies. *metaFARVAT* can also be applied to quantitative phenotypes and is able to combine all study designs, including family-based, case-control, and population-based designs. Furthermore, the proposed method was applied to a meta-analysis of EOCOPD and COPDGene data, and *DLEC1* was found to be genome-wide significant. *DLEC1* is a protein-coding gene encoding a cilia and flagella-associated protein. This gene has been implicated in several cancers but has not been previously associated with COPD. However, cilia-associated genes have been previously implicated in COPD (Tilley et al. 2015).

Despite the robustness and efficiency of the proposed method, there are still some limitations of the developed method. First, VT methods sort rare variants according to their MAFs and search the optimal threshold for rare variants. This approach is useful when it is not clear how to define rare variants. However, I found that TIE can be inflated if the number of rare variants is too small, and it is computationally intensive if there are a large number of variants to investigate. This problem can be solved by using a permutation method, and further investigation of this approach is necessary. Secondly, sufficiently large samples are necessary to guarantee that SKAT-O follows the assumed asymptotic distribution of the SKAT-O approach under the null hypothesis. Therefore, the SKAT-O type *metaFARVAT* also has this limitation when it is applied to a dichotomous phenotype with a small sample size. Thirdly, the proposed method cannot be applied to X- or Y-linked genes because the distributions of variants in X and Y chromosomes are different in males and females. Such an improvement will be considered in our future work. Fourthly, *metaFARVAT* requires raw data, which includes phenotype and genotype from each study. It cannot be directly applied to summary statistics from studies, such as score statistic and its variance, unless the statistics are generated using *FARVAT*. Lastly, in the simulation studies, I considered a limited number of rare variants and excluded noise variants. However, in practice, it is not known which rare variants are causal and which represent noise. Extensive simulations are thus necessary in our future work.

Despite the importance of rare variant analyses with family-based samples, this field of study has suffered over the last decades from a lack of statistical methods. In this study, I proposed new methods for family-based samples, enabling such statistical analyses.

Chapter 5

Summary & Conclusions

In spite of the success of GWAS in discovering DSL, it only identified a limited number of loci that partially explain disease heritability. Sequencing technology was expected to supply this additional information by obtaining large stretches of DNA spanning the entire genome, and improvements in this technology have enabled genetic association analysis of rare/common causal variants. Several rare variant association methods have been proposed. However, due to genetic heterogeneity and small sample size, very few genome-wide significant results have been found. In this thesis, I focused on the approaches that can enrich genetic effects and improve statistical power of rare variant association tests.

In chapter 2, I overviewed family-based association studies and compared the existing family-based rare variant association tests with GAW19 data. I found *FARVAT* is the most robust, statistically powerful, and computationally efficient method. Therefore, I extended *FARVAT* to multiple phenotype analysis and meta-analysis which were described in chapter 3 and chapter 4.

In chapter 3, I propose a new method for family-based rare variants associated with dichotomous phenotypes, quantitative phenotypes, or both. The proposed method enables multivariate analyses of extended families to detect rare variants under homogeneous and heterogeneous disease models. Extensive simulation studies show that *mFARVAT* works well for dichotomous and quantitative phenotypes. Our method is computationally efficient and association analyses at the genome-wide scale are computationally feasible for extended families. In our analyses, an Intel (R) Xeon (R) E5-2620 0 CPU at 2.00GHz, with a single node and 80 gigabyte memory, required six minutes to analyze the real data on two phenotypes. *mFARVAT* is implemented in C++.

In chapter 4, I proposed a novel meta-analysis method for family-based rare variant association analyses with both dichotomous phenotypes, which can test both homogeneous and heterogeneous effects of variants in different studies. *metaFARVAT* can also be applied to quantitative phenotypes and is able to combine all study designs, including family-based, case-control, and population-based designs. Furthermore, the proposed method was applied to a meta-analysis of EOCOPD and COPDGene data, and *DLEC1* was found to be genome-wide significant. *DLEC1* is a protein-coding gene encoding a cilia and

flagella-associated protein. This gene has been implicated in several cancers but has not been previously associated with COPD. However, cilia-associated genes have been previously implicated in COPD (Tilley et al. 2015).

In summary, family-based rare variant association tests with extension of multiple phenotype analysis and meta-analysis can overcome the limitations of traditional rare variant analysis, significantly improve statistical power, and reduce false-positive results. The proposed methods can be applied to various types of data, including population- and family-based designs, dichotomous and quantitative phenotypes, and homogenous and heterogeneous disease models. Furthermore, the combination of *mFARVAT* and *metaFARVAT* would be considered a good strategy to efficiently enrich genetic effects and identify trait- and disease-associate rare variants.

Future Work

We aim to build an all-in-one tool for family-based rare variant association studies. So far, *FARVAT* and its extension can analysis autosome and sex chromosomes (*FARVATX*) (Choi et al. 2017), dichotomous and quantitative phenotype, multiple phenotypes (*mFARVAT*), meta-analysis (*metaFARVAT*). There always been needs to identify rare variants associated with time-to-event traits, such as, age at disease onset, time to mortality, or time to secondary complications of disease. However, the existing methods of these association tests for family designs are limited. Therefore, my next study will extend FARVAT to survival traits (*sFARVAT*). I will derive an SKAT-O score for time-to-event outcomes in a Cox proportional hazard model framework and will

adapt a small adjustment procedure based on a higher moments matching method (Zhou et al. 2018) when analytical p-values are conservative. Li et al. (Li et al. 2019) provided a novel dynamic scan-statistic method, SCANG, which flexibly detects the sizes and the locations of rare variant association regions without the need to specify region set. This method can be potentially adapted to *sFARVAT*. Moreover, I will extend it to sex chromosome as *sFARVATX*.

Bibliography

Aaij, R., C. Abellan Beteta, A. Adametz, B. Adeva, M. Adinolfi, C. Adrover, et al. (2013). "Measurement of the Lambda(b)(0), Xi (b)(-), and Omega(b((-) Baryon masses." Phys Rev Lett **110**(18): 182001.

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, et al. (2010). "A method and server for predicting damaging missense mutations." Nat Methods **7**(4): 248-249.

Asimit, J. and E. Zeggini (2010). "Rare variant association analysis methods for complex traits." Annu Rev Genet **44**: 293-308.

Bourgain, C., S. Hoffjan, R. Nicolae, D. Newman, L. Steiner, K. Walker, et al. (2003). "Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus." Am J Hum Genet **73**(3): 612-626.

Chen, H., T. Lumley, J. Brody, N. L. Heard-Costa, C. S. Fox, L. A. Cupples, et al. (2014). "Sequence kernel association test for survival traits." Genet Epidemiol **38**(3): 191-197.

Chen, M. H. and Q. Yang (2010). "GWAF: an R package for genome-wide association analyses with family data." Bioinformatics **26**(4): 580-581.

Choi, S., S. Lee, S. Cichon, M. M. Nothen, C. Lange, T. Park, et al. (2014). "FARVAT: a family-based rare variant association test." Bioinformatics **30**(22): 3197-3205.

Choi, S., S. Lee, S. Cichon, M. M. Nothen, C. Lange, T. Park, et al. (2014). "FARVAT: a family-based rare variant association test." Bioinformatics.

Choi, S., S. Lee, Qiao, D., Hardin, M., Cho, MH., Silverman, EK., Park, T. Won, S. (2017) "FARVATX: FAmily-based Rare Variant Association Test for X-linked genes." Genet Epidemiol **40**(6): 475–485.

Published in final edited form as: Genet Epidemiol. 2016 Sep; 40(6): 475–485. Published online 2016 Jun 21. doi: 10.1002/gepi.21979

Choi, Y., E. M. Wijsman and B. S. Weir (2009). "Case-control association testing in the presence of unknown relationships." Genet Epidemiol **33**(8): 668-678.

Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, et al. (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly (Austin) **6**(2): 80-92.

Cirulli, E. T. and D. B. Goldstein (2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing." Nat Rev Genet **11**(6): 415-425.

Crowder, M. (1985). "Gaussian Estimation for correlated binomial data." Journal of the Royal Statistical Society B **1985**: 229–237.

Crowder, M. (1987). "On linear and quadratic estimating functions." Biometrika **74**: 591–597.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, et al. (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158.

Derkach, A., J. F. Lawless and L. Sun (2013). "Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests." Genet Epidemiol **37**(1): 110-121.

Derkacheva, M. and L. Hennig (2014). "Variations on a theme: Polycomb group proteins in plants." J Exp Bot **65**(10): 2769-2784.

Engelman, C. D. Greenwood, C. M., T. Bailey, J. N. Cantor, R. M. et al. (2016). "Genetic Analysis Workshop 19: methods and strategies for analyzing human sequence and gene expression data in extended families and unrelated individuals." BMC Proceedings **10**(7): 19.

Eichler, E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." Nat Rev Genet **11**(6): 446-450.

Feng, S., D. Liu, X. Zhan, M. K. Wing and G. R. Abecasis (2014). "RAREMETAL: fast and powerful meta-analysis for rare variants." Bioinformatics **30**(19): 2828-2829.

Genomes Project, C., G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, et al. (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature **491**(7422): 56-65.

George, V. T. and R. C. Elston (1987). "Testing the association between polymorphic markers and quantitative traits in pedigrees." Genet Epidemiol **4**(3): 193-201.

Gibson, G. (2012). "Rare and common variants: twenty arguments." Nat Rev Genet **13**(2): 135-145.

Gilmour, A. R., R. D. Anderson and A. Rae (1985). "The analysis of binomial data by a generalized linear mixed model." Biometrika **72**: 539–599.

Goldin, L. R., M. M. Martinez and E. S. Gershon (1991). "Sampling strategies for linkage studies." Eur Arch Psychiatry Clin Neurosci **240**(3): 182-187.

Gray-McGuire, C., M. Bochud, R. Goodloe and R. C. Elston (2009). "Genetic association tests: a method for the joint analysis of family and case-control data." Hum Genomics **4**(1): 2-20.

Han, F. and W. Pan (2010). "A data-adaptive sum test for disease association with multiple common or rare variants." Hum Hered **70**(1): 42-54.

He, Z., Zhang, D., Renton, A. E., Li, B., Zhao, L., Wang, G. T., et al. (2017). "The rare-variant generalized disequilibrium test for association analysis of nuclear and extended pedigrees with application to Alzheimer disease WGS data." Am. J. Hum. Genet **100**: 193–204.

Hoffmann, T. J., N. J. Marini and J. S. Witte (2010). "Comprehensive approach to analyzing rare genetic variants." PLoS One **5**(11): e13584.

Laird, N. M., S. Horvath and X. Xu (2000). "Implementing a unified approach to family-based tests of association." Genet Epidemiol **19 Suppl 1**: S36-42.

Laird, N. M. and C. Lange (2006). "Family-based designs in the age of large-scale gene-association studies." Nat Rev Genet **7**(5): 385-394.

Lange, C., D. L. DeMeo and N. M. Laird (2002). "Power and design considerations for a general class of family-based association tests: quantitative traits." Am J Hum Genet **71**(6): 1330-1341.

Lee, S., M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, et al. (2012). "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies." Am J Hum Genet **91**(2): 224-237.

Lee, S., T. M. Teslovich, M. Boehnke and X. Lin (2013). "General framework for meta-analysis of rare variants in sequencing association studies." Am J Hum Genet **93**(1): 42-53.

Lee, S., M. C. Wu and X. Lin (2012). "Optimal tests for rare variant effects in sequencing association studies." Biostatistics **13**(4): 762-775.

Lee, Y., S. Park, S. Moon, J. Lee, R. C. Elston, W. Lee, et al. (2014). "On the analysis of a repeated measure design in genome-wide association analysis." Int J Environ Res Public Health **11**(12): 12283-12303.

Li, B. and S. M. Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." Am J Hum Genet **83**(3): 311-321.

Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Morrison, AC., Boerwinkle, E., Lin, X.. (2019) "Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole Genome Sequencing Studies." Am J Hum Genet **104**(5): 802-814.

Lin, D. Y. and Z. Z. Tang (2011). "A general framework for detecting disease associations with rare variants in sequencing studies." Am J Hum Genet **89**(3): 354-367.

Liu, D. J., G. M. Peloso, X. Zhan, O. L. Holmen, M. Zawistowski, S. Feng, et al. (2014). "Meta-analysis of gene-level tests for rare variant association." Nat Genet **46**(2): 200-204.

Ma, C., T. Blackwell, M. Boehnke, L. J. Scott and T. D. i. Go (2013). "Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants." Genet Epidemiol **37**(6): 539-550.

MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, et al. (2012). "A systematic survey of loss-of-function variants in human protein-coding genes." Science **335**(6070): 823-828.

Madsen, B. E. and S. R. Browning (2009). "A groupwise association test for rare mutations using a weighted sum statistic." PLoS Genet **5**(2): e1000384.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, et al. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

McPeek, M. S., X. Wu and C. Ober (2004). "Best linear unbiased allele-frequency estimation in complex pedigrees." Biometrics **60**(2): 359-367.

Merikangas, K. R., M. A. Spence and D. J. Kupfer (1989). "Linkage studies of bipolar disorder: methodologic and analytic issues. Report of MacArthur Foundation Workshop on Linkage and Clinical Features in Affective Disorders." Arch Gen Psychiatry **46**(12): 1137-1141.

Moon, S., Y. Lee, S. Won and J. Lee (2018). "Multiple genotype-phenotype association study reveals intronic variant pair on SIDT2 associated with metabolic syndrome in a Korean population." Hum Genomics **12**(1): 48.

Morel, J. G., M. C. Bokossa and N. K. Neerchal (2003). "Small Sample Correction for the Variance of GEE Estimators." Biometrical Journal **4**: 395-409.

Morgenthaler, S. and W. G. Thilly (2007). "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST)." Mutat Res **615**(1-2): 28-56.

Morris, A. P., B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segre, V. Steinthorsdottir, et al. (2012). "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes." Nat Genet **44**(9): 981-990.

Morris, A. P. and E. Zeggini (2010). "An evaluation of statistical approaches to rare variant analysis in genetic association studies." Genet Epidemiol **34**(2): 188-193.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, et al. (2011). "Testing for an unusual distribution of rare variants." PLoS Genet **7**(3): e1001322.

Nejentsev, S., N. Walker, D. Riches, M. Egholm and J. A. Todd (2009). "Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes." Science **324**(5925): 387-389.

Pan, W. (2009). "Asymptotic tests of association with multiple SNPs in linkage disequilibrium." Genet Epidemiol **33**(6): 497-507.

Price, A. L., G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples, L. J. Wei, et al. (2010). "Pooled association tests for rare variants in exon-resequencing studies." Am J Hum Genet **86**(6): 832-838.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-575.

Qiao, D., C. Lange, T. H. Beaty, J. D. Crapo, K. C. Barnes, M. Bamshad, et al. (2016). "Exome Sequencing Analysis in Severe, Early-Onset Chronic Obstructive Pulmonary Disease." Am J Respir Crit Care Med **193**(12): 1353-1363.

Qiao, D. D., C. Lange, T. H. Beaty, J. D. Crapo, K. C. Bames, M. Bamshad, et al. (2016). "Exome Sequencing Analysis in Severe, Early-Onset Chronic Obstructive Pulmonary Disease." American Journal of Respiratory and Critical Care Medicine **193**(12): 1353-1363.

Qiao, D. D., Ameli, A., Prokopenko, D., Chen, H., et al. (2018). "Whole exome sequencing analysis in severe chronic obstructive pulmonary disease." Human Molecular Genetics **27**(21): 3801–3812.

Regan, E. A., J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, et al. (2010). "Genetic epidemiology of COPD (COPDGene) study design." COPD **7**(1): 32-43.

Sasieni, P.D. (1997). "From genotypes to genes: doubling thesample size." **Biometric**s **53**: 1253–1261.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly and D. Altshuler (2005). "Calibrating a coalescent simulation of human genome sequence variation." Genome Res **15**(11): 1576-1583.

Schaid, D. J., S. K. McDonnell, J. P. Sinnwell and S. N. Thibodeau (2013). "Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data." Genet Epidemiol **37**(5): 409-418.

Schall, R. (1991). "Estimation in generalized linear models with random effects." Biometrika **78**: 719–727.

Schifano, E. D., L. Li, D. C. Christiani and X. Lin (2013). "Genome-wide association analysis for multiple continuous secondary phenotypes." Am J Hum Genet **92**(5): 744-759.

Shi, G. and D. C. Rao (2011). "Optimum designs for next-generation sequencing to discover rare variants for common complex disease." Genet Epidemiol **35**(6): 572-579.

Silverman, E. K., H. A. Chapman, J. M. Drazen, S. T. Weiss, B. Rosner, E. J. Campbell, et al. (1998). "Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis." Am J Respir Crit Care Med **157**(6 Pt 1): 1770-1778.

Slager, S. L. and D. J. Schaid (2001). "Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects." Am J Hum Genet **68**(6): 1457-1462.

Spielman, R. S., R. E. McGinnis and W. J. Ewens (1993). "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)." Am J Hum Genet **52**(3): 506-516.

Sun, J., Y. Zheng and L. Hsu (2013). "A unified mixed-effects model for rare-variant association in sequencing studies." Genet Epidemiol **37**(4): 334-344.

Sun, L., K. Wilder and M. S. McPeek (2002). "Enhanced pedigree error detection." Hum Hered **54**(2): 99-110.

Tang, Z. Z. and D. Y. Lin (2013). "MASS: meta-analysis of score statistics for sequencing studies." Bioinformatics **29**(14): 1803-1805.

Tang, Z. Z. and D. Y. Lin (2014). "Meta-analysis of sequencing studies with heterogeneous genetic associations." Genet Epidemiol **38**(5): 389-401.

Tang, Z. Z. and D. Y. Lin (2015). "Meta-analysis for Discovering Rare-Variant Associations: Statistical Methods and Software Programs." Am J Hum Genet **97**(1): 35-53.

Thornton, T. and M. S. McPeek (2007). "Case-control association testing with related individuals: a more powerful quasi-likelihood score test." Am J Hum Genet **81**(2): 321-337.

Thornton, T. and M. S. McPeek (2010). "ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure." Am J Hum Genet **86**(2): 172-184.

Thornton, T., H. Tang, T. J. Hoffmann, H. M. Ochs-Balcom, B. J. Caan and N. Risch (2012). "Estimating kinship in admixed populations." Am J Hum Genet **91**(1): 122-138.

Tilley, A. E., M. S. Walters, R. Shaykhiev and R. G. Crystal (2015). "Cilia dysfunction in lung disease." Annu Rev Physiol **77**: 379-406.

van der Sluis, S., D. Posthuma and C. V. Dolan (2013). "TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies." PLoS Genet **9**(1): e1003235.

Wang, X., S. Lee, X. Zhu, S. Redline and X. Lin (2013). "GEE-based SNP set association test for continuous and discrete traits in family-based association studies." Genet Epidemiol **37**(8): 778-786.

Weir, B. S., A. D. Anderson and A. B. Hepler (2006). "Genetic relatedness analysis: modern data and new challenges." Nat Rev Genet **7**(10): 771-780.

Won, S., W. Kim, S. Lee, Y. Lee, J. Sung and T. Park (2015). "Family-based association analysis: a fast and efficient method of multivariate association analysis with multiple variants." BMC Bioinformatics **16**: 46.

Won, S. and C. Lange (2013). "A general framework for robust and efficient association analysis in family-based designs: quantitative and dichotomous phenotypes." Stat Med **32**(25): 4482-4498.

Won, S. and C. Lange (2013). "A general framework for robust and efficient association analysis in family-based designs: quantitative and dichotomous phenotypes." Stat Med.

Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, et al. (2010). "Powerful SNP-set analysis for case-control genome-wide association studies." Am J Hum Genet **86**(6): 929-942.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke and X. Lin (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." Am J Hum Genet **89**(1): 82-93.

Xiong, M., J. Zhao and E. Boerwinkle (2002). "Generalized T2 test for genome association studies." Am J Hum Genet **70**(5): 1257-1268.

Yan, Q., Tiwari, H. K., Yi, N., Gao, G., Zhang, K., Lin, W. Y., et al. (2015). "A sequence kernel association test for dichotomous traits in family samples under a generalized linear mixed model." Hum. Hered **79**: 60–68.

Yip, W., G. De, A. B. Raby and N. Laird (2011). "Identifying causal rare variants of disease through family-based analysis of Genetics Analysis Workshop 17 data set." BMC Proceedings.

Zhou, W., Nielsen, JB., Fritsche, LG., Dey, R., Gabrielsen, ME., Wolford, BN., LeFaive, J., VandeHaar, P., Gagliano, SA., Gifford, A., Bastarache, LA., Wei, WQ., Denny, JC., Lin, M., Hveem, K., Kang, HM., Abecasis, GR., Willer, CJ., Lee, S. (2018) "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies." Nat Genet **50**(9): 1335–1341.

Zhu, Y. and M. Xiong (2012). "Family-based association studies for next-generation sequencing." Am J Hum Genet **90**(6): 1028-1045.

초 록

수많은 전장유전체연관분석(GWAS)에도 불구하고 질병연관 유전체변이(DSL)는 제한적으로만 발견되었는데 이는 실종된 질병유전성(missing heritability)에 기인한다. 한 번에 긴 리드(read)를 시퀀싱하는 기술은 이를 보완해 줄 것으로 기대되어 왔으며, 이 기술의 발달 덕분에 유전체연관분석을 활용하여 여러 희귀(rare) 및 일반(common) 인과 변이를 발견할 수 있었다. 그러나 꽤 많은 샘플을 이용한 실험에서도 단일 변이를 대상으로한 전장유전체연관분석은 부정오류(false negative) 문제에서 자유로울 수 없다. 이에 희귀변이 연관 분석의 검정력을 증가시키기 위해 생물학적으로 연관이 있는 위치의 여러 유전체변이를 하나로 합쳐서 분석하는 방법들이 제안되었다. 버든 검정(burden test), 분산구조 검정(variance component test), 결합 옴니버스 검정(combined omnibus test) 등의 위치기반 연관 분석이 바로 그것이다.

희귀변이 연관분석에 위와 같은 분석방법을 활용하면 검정력이 크게 증가하여 더 많은 질병연관 유전체 변이를 발견할 수 있을 것으로 기대되어왔다. 하지만 샘플 간 유전적 이질성의 존재와 상대적으로 샘플 수가 적은 한계들 때문에 매우 적은 수의 변이만이 발견되었다. 이러한 문제점을 해결하기 위해 다양한 방법들이 개발되었는데, 그 중 하나는 가족기반 분석 방법으로 이는 샘플 간 유전적 이질성과 집단층화 문제를 다루는데 용이하다. 두 번째로 서로 다른 표현형이 서로 관련이 있을 경우 검정력을 증가시키기 위해 이들을 한번에 분석하는 방법이 있다. 세 번째는 메타분석을 활용하여 여러 연구의 결과를 합치는 방법으로 이는 많은 연구들에서 효과적임이 밝혀졌다.

이 논문에서는 현재 많이 사용되고 있는 여러 가족기반 희귀변이 연관 분석 방법을 비교하였고 다른 방법들에 비해 FARVAT 이 통계적으로 견고하며 계산 효율적인 방법임을 보였다. 더 나아가 이를 다중 표현형 분석 방법(*mFARVAT*)과 메타분석

방법(*metaFARVAT*)으로 확장하였다. *mFARVAT*은 유사우도함수 기반 스코어 테스트(quasi-likelihood-based score test)를 다수의 표현형에 적용하는 희귀질환 연관분석 방법으로 표현형들에 대한 각 변이의 동질성 및 이질성 효과를 검증한다. *metaFARVAT* 은 여러 연구에서의 유도함수 스코어를 결합하여 버든 통계량, 변이 임계(variable threshold) 통계량, 분산구조 통계량, 결합 옴니버스 통계량을 생성한다. 이는 여러 연구들의 결과를 이용하여 변이들의 동질성 및 이질성 효과를 검증하며, 정량 표현형 및 이분 표현형에 적용이 가능하다. 다양한 시나리오 하에서의 광범위한 모의 실험을 통해 제안한 방법들이 일반적으로 견고하고 효율적이라는 것을 보였다. 또한 이 방법을 활용하여 DLEC1 등의 만성폐쇄성폐질환(COPD) 관련 후보 유전자를 발견하였다.

주요어: 희귀변이 연관 분석, 가족 기반 분석, 다중 표현형, 메타 분석, 만성폐쇄성폐질환

학 번: 2015-30742