이 학 박 사 학 위 논 문

Statistical Analysis for Next-Generation Sequencing data in Family-based designs

가족 기반 차세대 염기서열자료의 통계적 분석 연구

2016 년 8 월

서울대학교 대학원

협동과정 생물정보학과

최 성 경

Statistical Analysis for Next-Generation Sequencing data in Family-based designs

by

Sungkyoung Choi

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, 2016

Statistical Analysis for Next-Generation Sequencing data in Family-based designs

지도교수 박 태 성

이 논문을 이학박사 학위논문으로 제출함

2016 년 **8** 월

서울대학교 대학원

생물정보협동과정 생물정보학 전공

최 성 경

최성경의 이학박사 학위논문을 인준함

2016 년 **8** 월

위 원 장 유 연 주 (인)

부위원장 박 태 성 (인)

위 원 성 주 헌 (인)

위 원 이 승 연 (인)

위 원 원 성 호 (인)

Abstract

Statistical Analysis for Next-Generation Sequencing data in Family-based designs

Sungkyoung Choi

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

Genome-wide association studies (GWAS) typically involve examination of 100,000 to more than 1,000,000 genetic variants, such as single nucleotide polymorphisms (SNPs), in different individuals to identify SNPs associated with a disease. Since the conclusion of the Human Genome Project, this project elucidated understand human genetic variation and paved the way for the GWAS approach. GWAS have successfully identified thousands of common genetic loci associated with many phenotypes.

Despite the success of GWAS, the variants identified by these studies have often explained only a small fraction of heritability for most phenotypes, and this observation underscored the importance of studying rare or less common variants.

Contrary to the traditional GWAS approach, single variant association analysis with rare variants has difficulties with detection of causal variants. To overcome the issue with statistical power in rare variant association studies, researchers have recently developed statistical methods for testing rare variants in a population-based design. Because individuals in a family are genetically more homogenous than unrelated individuals, family-based designs can play an important role in rare-variant studies. Despite the importance of rare variant analysis for the family-based design, only a few statistical methods for family-based rare-variant association analysis are available. Furthermore, even though many genes on the X chromosome are related to human complex diseases, few significantly associated rare variants have been identified on the X chromosome for complex traits.

In this study, we propose a family-based rare-variant association test (***FARVAT***) and a family-based rare-variant association test for X-linked genes (***FARVATX***). ***FARVAT*** is based on quasi-likelihood, and is statistically and computationally efficient for the family-based design. We considered that families were ascertained with the disease status of family members, and we calculated the genetic relationship matrix for the proposed method; this matrix ensured robustness in the presence of population substructure. Depending on the choice of a working matrix, ***FARVAT*** could be a burden-type or a variance component-type method, and could be extended to the optimal-type method. In the analysis of the X chromosome, ***FARVATX*** can accommodate random X chromosome inactivation (XCI), escaped XCI (E-XCI), and skewed XCI (S-

XCI). ***FARVATX*** is computationally efficient and can complete X-linked analyses within a few hours. With extensive simulation studies under various scenarios, we compared the proposed methods with the existing ones, and the results showed that the proposed methods are the more powerful in terms of simulation settings. We also applied ***FARVAT*** and ***FARVATX*** to schizophrenia data and chronic obstructive pulmonary disease (COPD) data, respectively. The proposed methods may help researchers identify additional X-linked rare variants associated with complex traits, thereby leading to a better understanding of the underlying biological processes associated with X-linked genes.

Key words: Genome-wide association studies (GWAS), Next-Generation Sequencing (NGS), rare-variant association test, Family-based design, X chromosome

Student number: 2011-30927

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The background on genetic association studies

1.1.1 An overview of genome-wide association studies

A genome-wide association study (GWAS) represents an approach to identify causal variants that are associated with complex traits in the population [Visscher, et al. 2012]. The causal variant is a single nucleotide polymorphism (SNP) that is known to be highly associated with increased or decreased individual risk of a disease. A SNP is a genetic variation when a single-nucleotide (adenine, guanine, cytosine, or thymine) variation in a segment of a DNA occurs in more than one percent of a population. Because a

SNP is generally bi-allelic, we considered three possible genotypes: *A*/*A*, *A*/*a*, and *a*/*a* (minor allele *a* and major allele *A*).

The GWAS approach was first proposed by researchers [Risch and Merikangas 1996] who claimed that association studies are more powerful than the linkage study design in terms of detection of common variants. Another researcher advanced the common-disease common variant (CDCV) hypothesis [Lander 1996] stating that common disorders are influenced by a genetic variation that is also common in the general population. A SNP becomes common over many generations because natural selection has filtered out any disease-causing mutations. Nonetheless, some common SNPs have harmful effects depending on environmental conditions. The CDCV-GWAS strategy implies that many common SNPs have small effects on each disease, and that some could be discovered by testing enough SNPs in a large sample of people [Psychiatric, et al. 2009].

The International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) has validated approximately 3.1 million SNPs in major populations [International HapMap, et al. 2007]. Companies Affymetrix and Illumina have rapidly developed SNP arrays with high accuracy at a low cost [Psychiatric, et al. 2009]. These efforts make it possible to capture most of the common genomic variation in a number of human populations using representative tag SNPs.

The GWAS has successfully identified more than 10,000 common genetic variants associated with many complex human traits [Altshuler, et al. 2008; Manolio, et al. 2008; McCarthy, et al. 2008]. Since the first major

GWAS was reported in 2007 [Sladek, et al. 2007], such studies have rapidly grown in scale and complexity, and 1,751 curated publications of 11,912 SNPs have been added to the catalog of published Genome-wide Association Studies [Welter, et al. 2014] (See Figure 1.1).

In a GWAS, the association analysis is mainly focused on autosomal chromosomes only, whereas those for the X chromosome is neglected. Due to the relatively large size of the X chromosome, many X-linked genes have important functions, and significant associations of several X-linked variants have been identified for diverse phenotypes, including blood pressure, hematological traits, obesity, HDL cholesterol, and Type-1 diabetes [Ahituv, et al. 2007; Auer, et al. 2014; Blakemore, et al. 2009; Cohen, et al. 2004; Gaukrodger, et al. 2005; Nejentsev, et al. 2009]. However, most successful results from genome-wide association studies have been from autosomes, and significant results for X-linked variants are relatively few in number. There are multiple potential reasons for this, but it is at least partially attributable to the complex biological properties of X-linked variants, which make efficient genetic association analyses more challenging. For instance, while females inherit X chromosomes from both parents, males inherit a single maternal X chromosome, and there is some empirical evidence that in females genes for some X-linked variants are expressed twice as highly as in males [Brown and Greally 2003; Carrel and Willard 2005; Shapiro, et al. 1979]. In contrast, dosage compensation for other X-linked variants can be achieved by the selection, and silencing of maternal or paternal genes via either random or

nonrandom mechanisms [Lyon 1961]. Under nonrandom X chromosome inactivation (XCI), either the maternal or paternal genes are relatively more activated [Belmont 1996; Plenge, et al. 2002], and the amount of skewness is sometimes related to age or disease status [Amos-Landgraf, et al. 2006; Busque, et al. 1996; Chagnon, et al. 2005; Knudsen, et al. 2007; Minks, et al. 2008; Sharp, et al. 2000; Wong, et al. 2011]. However in spite of this knowledge about gene expression process of X-linked variants, there are very few statistical methods applicable to the complicated biological process of X-linked genes.

1.1.1.1 The case-control design

In a GWAS, case-control status is generally analyzed using either contingency table methods or logistic regression. Contingency table methods test the null hypothesis of no association between rows and columns in a 2×3 matrix where the rows correspond to case-control status and the columns to three genotypes. Four genetic models named recessive (REC), dominant (DOM), additive (ADD), and multiplicative (MUL) models are commonly used [Freidlin, et al. 2002; Sasieni 1997] (See Table 1.1). Researchers can use either the chi-square test (with two degrees of freedom) or Fisher's exact test.

The Cochran-Armitage test [Armitage 1955], which is more conservative than the chi-square test and does not depend on the Hardy-Weinberg equilibrium, is similar to the allele-count test. The idea is to test the

hypothesis of a zero slope for a line that fits the three-genotype risk estimates best. Under the null hypothesis, the Cochran-Armitage test has an approximate the chi-square distribution with a single degree of freedom. The Cochran-Armitage test has good power in the additive model, but power is reduced by a deviation from additivity.

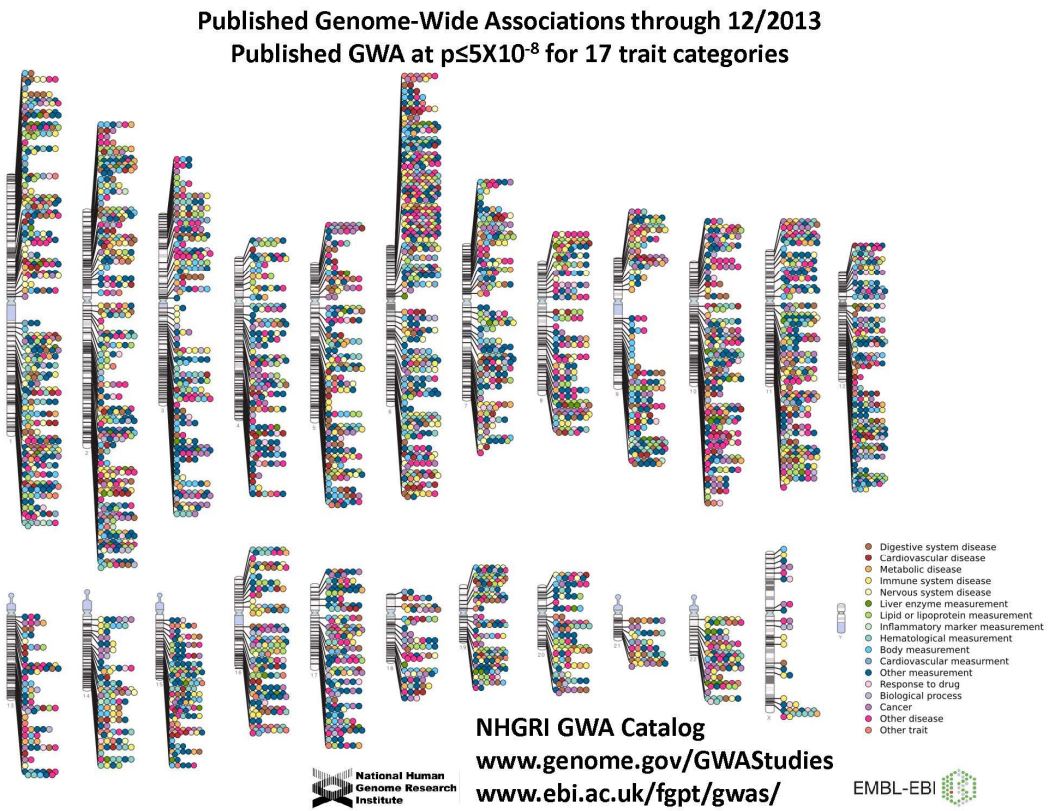


Figure 1.1 GWAS catalog as of 2013. This diagram shows all SNP-trait association with p-value $\leq 5.0\times10^{-8}$.

Table 1.1 Contingency tables for case-control analyses by genetic model. *a*, *b*, *c*, *d*, *e*, *f* are genotype count observed in cases and controls.

(a) Recessive model

	AA + AB	**BB**
Cases	*a* + *b*	*c*
Controls	*d* + *e*	*f*

(b) Dominant model

	AA	**AB+BB**
Cases	*a*	*b* + *c*
Controls	*d*	*e* + *f*

(c) Additive model

	AA	**AB**	**BB**
Cases	*a*	*b*	*c*
Controls	*d*	*e*	*f*

(d) Multiplicative model

	A	**B**
Cases	2*a* + *b*	*b* + 2*c*
Controls	2*d* + *e*	*e* + 2*f*

Logistic regression is an extension of linear regression where the outcome of a linear model is transformed using a logistic function, $\text{logit}(\pi)=\log(\pi/(1-\pi))$, that predicts the probability of having a case status given a genotype class. Logistic regression is preferred for estimates of genetic risk factors because it allows for adjustment for covariates such as age, sex, and many others that may influence the phenotype.

Several methods have been proposed to detect statistically associated X-linked variants of phenotypes. For X-linked variants, there is heterogeneity of genetic distributions between males and females, which is often handled by extending the Cochran-Armitage test for genetic association analyses of X-linked variants [Clayton 2008; Zheng, et al. 2007].

1.1.1.2 The family-based design

The family-based designs have an advantage over case-control designs because the former are robust in the presence of a population admixture and stratification. The simplest family-based design for testing of an association uses genotype data from a trio, which consists of an affected offspring and two parents. The association in this design is tested using the transmission disequilibrium test (TDT) [Spielman, et al. 1993]. The TDT is focused on heterozygous parents and tests whether a specific variant has an equal frequency among the variants inherited and those not inherited by the affected offspring. Originally, the TDT was proposed as a test of linkage in the presence of an association. Because both linkage and association between the

trait and the variant have to be present for the TDT to reject the null hypothesis [Ott 1989], the TDT is now typically used to test for an association [Hirschhorn and Daly 2005]. The TDT does not require specification of a disease model. Nevertheless, there are many cases where the original TDT cannot be applied without extension, for instance, missing parents, general pedigrees, and complex phenotypes.

A general extension of the TDT has resulted in the family-based association test (FBAT) approach [Laird and Lange 2006]. The FBAT approach incorporates additional conditions such as general pedigrees, missing founders, and complex diseases. The FBAT approach has been successfully applied, for instance, to studies of asthma [Smit, et al. 2009].

In order to analyze the X chromosome, some researchers [Thornton, et al. 2012] proposed methods, X_M, in case-control study with family-based samples. The X_M method adjusts for both relationships among family members and sex-specific trait prevalence values.

1.1.2 An overview of next-generation sequencing data analysis

As the cost of genotyping decreases, the number of the GWAS has increased substantially, and the GWAS approach is now relatively common. Despite the success of the GWAS at identifying common SNPs that are associated with complex diseases [Manolio, et al. 2008; Visscher, et al. 2012], heritability of complex traits is only partially explained by these significant variants from the GWAS [Manolio, et al. 2009]. For example, it is estimated that the heritability of human height is ~80%, but 40 loci that have been associated with height, explain only ~5% of height variance [Visscher 2008]. This is known as the problem of missing heritability [Manolio, et al. 2009], and one possible solution to this problem, the analysis of rare variants, is generally not feasible for the GWAS [Li and Leal 2008; Wu, et al. 2011].

With the rapid development of cost-effective NGS technologies such as Illumina HiSeq, ABI SOLiD, and Roche 454, rare variant association analysis using sequence data has become possible. A number of studies provide evidence that rare variants contribute to the etiology of complex traits such as high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL), blood pressure, and triglyceride levels [Cohen, et al. 2004; Cohen, et al. 2006; Ji, et al. 2008; Romeo, et al. 2009]. Therefore, rare variants can play an important role in a complex disease.

1.1.2.1 The case-control design

Contrary to the analysis of common variants, single genetic association analysis with rare variants often yields a large percentage of false negatives unless sample sizes or effect sizes are very large. Accordingly, an association analysis with the collapsed genotype scores for a set of rare variants has been suggested [Li and Leal 2008]. For instance, minor alleles for all rare variants in a gene or a region are counted, and the disease status is regressed on minor allele counts (MAC). Alternatively, the collapsed amount of variance inflation for rare variants can be compared between affected and unaffected individuals [Neale, et al. 2011; Wu, et al. 2011]. The former is often called a burden test, while the latter a variance component test. The burden test is statistically more efficient than variance component methods such as C-alpha [Neale, et al. 2011] and sequence kernel test (SKAT) [Wu, et al. 2011] if most of the rare alleles have similar effects on the disease. On the other hand, if rare variants with deleterious and protective effects are combined, the collapsed genotype scores for affected and unaffected individuals are similar, and genetic association analysis with a burden test becomes inefficient, while the variance component method becomes more robust. The two methods can be combined into robust statistical strategies such as the SKAT-O approach [Lee, et al. 2012a], which is statistically efficient in both situations.

A research group [Ma, et al. 2015] proposed three gene-based test for the X chromosome: burden-type test, variance component-type test, and optimal-type tests.

1.1.2.2 The family-based design

Recently, some researchers [He, et al. 2014] proposed rare variant transmission disequilibrium tests (RV-TDTs) which represent extensions of the (TDT) [Spielman, et al. 1993]. The RV-TDT approach was shown to be robust and powerful for exploration of rare-variant associations in a population substructure. Even though robustness against the population substructure is ensured, these approaches do not take into account the parental phenotypes, and therefore power loss can be substantial for extended family designs. Alternatively, some investigators have proposed the family-based functional principal-component analysis (FPCA) and pedigree-based combined multivariate and collapsing (PedCMC) tests [Zhu and Xiong 2012], which are extended Cochran Armitage tests for family-based samples. These tests utilize data from the whole family for a rare variant association analysis and are expected to be more efficient than the TDT-type statistic. Nevertheless, if the effects of rare variants are proportional to MAC or if the protective and deleterious variants are mixed in a gene, these approaches can be less efficient.

Some researchers [Schaid, et al. 2013] proposed new methods for rare X-linked variant association analysis with family-based samples, which are modified the M_{QLS} method [Thornton and McPeek 2007]; the extensions are on the applicability from common X-linked variant to rare variant analysis on the X chromosome in the family-based design.

1.2 The purpose of this study

The main purpose of this thesis is to develop statistical methods for identification of rare causal variants using family-based sequencing data. To this end, the thesis is focused on two type of study. One is a study to develop a new method, which is based on the quasi-likelihood of whole families. To increase statistical power, the proposed method utilizes the disease prevalence information and the genetic relationship matrix. The other is intended to extend our previous study from autosomes to the X chromosome.

In the first study, we proposed a family-based rare variant association test (***FARVAT***) for NGS data. The previous approaches to detect human disease associations with rare variants in a family-based design are not powerful and cannot be considered in the presence of both deleterious and protective rare variants in a gene. To overcome the limitation of previous approaches, we developed the ***FARVAT*** method, which utilizes the prevalence of a disease and the empirical genetic relationship matrix between individuals. Furthermore, the ***FARVAT*** method can be a burden-type test, variance component-type test, or optimal-type test depending on the choice of the working matrix. We evaluated the performance of the proposed method in a simulation study and applied this method to a real NGS dataset from 36 trios with schizophrenia.

In the second study, we proposed a novel method (for a family-based rare-variant association test for X-linked genes, ***FARVATX***) for identification

of rare variants on the X chromosome in a family-based design. In order to analyze the X chromosome, previous methods have been proposed, which modified the M_{QLS} method. Even though the powerful approaches to detect X chromosome genetic association are ensured, these approaches cannot be considered in the presence of skewed X-chromosome inactivation patterns in female or the different proportions of males and females. To overcome these limitations, we proposed FARVATX method that is applicable in various biological models such as random X chromosome inactivation (XCI), escaped XCI (E-XCI), and skewed XCI (S-XCI). Furthermore, the proposed method used an allele frequency estimation for X chromosome marker in samples with related individuals. The proposed method shows better performance than those of existing methods, in our extensive simulations. The proposed method was applied to an association analysis of families with chronic obstructive pulmonary disease (COPD).

1.3 Outline of the thesis

This thesis is organized as follows. Chapter 1 is an introduction to this study with an overview of GWAS and NGS analysis for the case-control design and family-based design. Chapter 2 contains an overview of rare-variant analysis. Chapter 3 deals with the family-based rare-variant association test. Chapter 4 is about the family-based rare-variant association test for X-linked genes. Chapters 3 and 4 contain an introduction to the statistical method, simulation studies, and the application to real data. Finally, the summary and conclusions are presented in Chapter 5.

Chapter 2

An overview of rare-variant analysis

2.1 Challenges of rare-variant analysis

With the rapid advances in the development of NGS technologies, the cost of sequencing has decreased and provided an opportunity to study the role of rare variants in human complex traits. Nonetheless, the analysis of rare variants in NGS data poses substantial challenges.

First, it is necessary to use a very large sample size to detect causal rare variants [Gorlov, et al. 2008; Li and Leal 2008]. For example, at the odds ratio (OR) of 1.4, the required sample sizes to achieve the power of 80% are 54,000 and 540,000 when a minor allele frequency (MAF) is assumed to be 0.01 and 0.001, respectively [Lee, et al. 2014]. To solve this problem, many alternative strategies have been proposed such as target sequencing, exome sequencing,

imputation for low-depth whole-genome sequencing, and the family-based design. Second, the analysis of rare variants in sequencing data suffers from an increased multiple-testing burden. One-third of all sequencing variants have MAFs below 5%, and the distribution of MAFs is substantially skewed toward an excess of rare variants [Braverman, et al. 1995; Gibson 2011]. Therefore, the significance level may have to be more stringent because of a large number of rare variants. Finally, classical single-variant-based tests for rare variants are seriously underpowered as compared to common variants because statistical power depends on MAF. To address these issues, numerous methods for detecting associations with rare variants for complex traits have been designed.

2.2 Review of rare-variant methods

To detect a causal rare variant, aggregation tests evaluate cumulative effects of multiple genetic variants in a gene or region instead of testing each variant individually. Numerous methods have been proposed, and we broadly categorize these methods into four types: the burden-type tests, the variance component-type tests, the optimal-type tests, and other tests (See Table 2.1).

Table 2.1 Summary of statistical methods for rare variant association testing

Study Design	Type	Methods	References
Case-Control Design	Burden type tests	Cohort Allelic Sums Tests (CAST)	[Morgenthaler and Thilly 2007]
		Combined Multivariate Collapsing (CMC)	[Li and Leal 2008]
		Weighted Sum Statistic (WSS)	[Madsen and Browning 2009]
		Kernel Based Adaptive Cluster (KBAC)	[Liu and Leal 2010]
		Variable Threshold (VT)	[Price, et al. 2010]
		Burden test to analyze X-chromosome variants	[Ma, et al. 2015]
	Variance component tests	C-alpha test	[Neale, et al. 2011]
		Sequence Kernel Association Test (SKAT)	[Wu, et al. 2011]
		SKAT to analyze X-chromosome variants	[Ma, et al. 2015]
	Optimal type tests	SKAT-O	[Lee, et al. 2012a]
		Fisher's hybrid statistics	[Derkach, et al. 2013]
		Mixed effects Score Test (MiST)	[Sun, et al. 2013]
		SKAT-O to analyze X-chromosome variants	[Ma, et al. 2015]

	Other tests	Replication Based Test (RBT)	[Ionita-Laza, et al. 2011]
		Variant Annotation, Analysis & Search Tool (VAAST)	[Yandell, et al. 2011]
		Exponential-Combination (EC) test	[Chen, et al. 2012]
		Optimal Combination of Single-variant Tests (OCST)	[Sha and Zhang 2014]
Family-based Design	**Burden type tests**	Family-based functional Principal-Component Analysis (FPCA)	[Zhu and Xiong 2012]
		Pedigree-based Combined Multivariate and Collapsing (PedCMC)	[Zhu and Xiong 2012]
		Rare Variant Transmission Disequilibrium Test (RV-TDT)	[He, et al. 2014]
		PedCMC-Burden for the X chromosome	[Schaid, et al. 2013]
	Variance component tests	family-based SKAT (famSKAT)	[Chen, et al. 2013]
		PedCMC-Kernel for the X chromosome	[Schaid, et al. 2013]
	Optimal type test	Minimum *p*-value Optimized Nuisance parameter Score Test Extended to Relatives (MONSTER)	[Jiang and McPeek 2014]
	Other test	pedigree-VAAST (pVAAST)	[Hu, et al. 2014]

2.2.1 Burden-type methods

The burden-type test collapses information on multiple rare variants into a genetic burden score and tests for association between this score and a trait. Several burden type approaches such as cohort allelic sums tests (CAST) [Morgenthaler and Thilly 2007], combined multivariate collapsing (CMC) [Li and Leal 2008], weighted sum statistic (WSS) [Madsen and Browning 2009], kernel-based adaptive cluster (KBAC) [Liu and Leal 2010], and variable threshold (VT) [Price, et al. 2010] have been proposed. In the family-based design, FPCA [Zhu and Xiong 2012], PedCMC [Zhu and Xiong 2012], and RV-TDT [He, et al. 2014] have been proposed. For rare X-linked variant association analysis, extensions of Burden test [Ma, et al. 2015] and PedCMC-Burden [Schaid, et al. 2013] have been proposed in case-control design and family-based design, respectively.

In the CAST method [Morgenthaler and Thilly 2007], the genetic burden score is given by

$$C_i = I_{\{\sum_{j=1}^{k} G_{ij} > 0\}},$$

where G_{ij} denotes the number of rare variants of subject i at variant j in a group of k genetic variants. I_A is an indicator variable assuming the value of 1 when A is true. The association test involves chi-square test or Fisher's exact test of a contingency table. The CAST method does not account for covariates, cannot be used with a continuous phenotype, and does not consider weighting of rare variants.

An extension of the CAST method is to consider a combination of rare variants and common variants in a CMC method [Li and Leal 2008]. The CMC method collapses rare variants within MAF-based subgroups and evaluates the genetic effects of both collapsed rare variants and common variants using Hotelling's T-test. The WSS method [Madsen and Browning 2009] assumes that rarer variants have stronger effects. The genetic burden score of the WSS method is expressed as

$$C_i = \sum_{j=1}^{k} w_j G_{ij},$$

where the weight, w_j, is assumed to be $1/[\mathrm{MAF}_j(1-\mathrm{MAF}_j)]^{1/2}$. The sum of ranks of the genetic burden score in the case group is then used as a summary statistic to be compared with those in the control group using a permutation method.

The KBAC method [Liu and Leal 2010] classifies rare variants into groups depending on the pattern of rare variants. The KBAC test statistic is

$$T_{KBAC} = \left(\sum_{j=1}^{k} w_j \left(\frac{n_j^{case}}{n_{case}} - \frac{n_j^{cont}}{n_{cont}} \right) \right)^2,$$

where n_j^{case} and n_j^{cont} are the numbers of rare variants at variant j in cases and controls, respectively, with $n_j = n_j^{case} + n_j^{cont}$. The weight, w_j, is determined by a hyper-geometric kernel:

$$w_j = \sum_{l=1}^{n_j^{case}} \frac{\binom{n_j}{1}\binom{n_{case}+n_{cont}-n_j}{n_{case}-1}}{\binom{n_{case}+n_{cont}}{n_{case}}}.$$

The *p*-value of KBAC is calculated by a permutation method.

The VT method [Price, et al. 2010] implies that the MAFs of causal rare variants may be different from those of non-causal rare variants. For a given threshold τ, the genetic burden score of the VT method is denoted by

$$C_i = \sum_{j=1}^{k} w_j G_{ij}, \text{ where } w_j = \begin{cases} 1 & \text{if } \mathrm{MAF}_j \le \tau \\ 0 & \text{if } \mathrm{MAF}_j > \tau \end{cases}.$$

The VT test statistic is

$$Z_{max} = max_{\tau} Z(\tau),$$

where $Z(t)$ is a z-score of C_i, and the *p*-value is calculated by a permutation method.

In a family-based study, the FPCA [Zhu and Xiong 2012] test statistic is

$$T_{\mathrm{FPCA}} = \frac{\left[\frac{n_{case}(n-n_{case})}{n}\right]^2 (\bar{\xi}_{case} - \bar{\xi}_{cont})^{\mathrm{T}} \Sigma_{\mathrm{FPCA}}^{-1} (\bar{\xi}_{case} - \bar{\xi}_{cont})}{\left(D_r - \frac{n_{case}}{n} D_p\right)^{\mathrm{T}} \mathbf{\Phi} \left(D_r - \frac{n_{case}}{n} D_p\right)},$$

where $\bar{\xi}_{\mathrm{case}}$ and $\bar{\xi}_{\mathrm{cont}}$ are the vector of averages of the functional principal-component scores in the case group and control group, respectively. D_r is defined as $[u_1, ..., u_n]^{\mathrm{T}}$, a column vector of length n, where

$$u_i = \begin{cases} 1 & \text{if } i \text{ is a case} \\ 0 & \text{if } i \text{ is a control} \end{cases}.$$

D_p is the $n\times1$ column vector that consists of 1s. $\mathbf{\Phi}$ is the kinship matrix. Σ_{FPCA} is defined as

$$\Sigma_{FPCA}=\begin{bmatrix}\sigma^{\xi}_{11} & \cdots & \sigma^{\xi}_{1k}\\ \vdots & \ddots & \vdots\\ \sigma^{\xi}_{k1} & \cdots & \sigma^{\xi}_{kk}\end{bmatrix},$$

where σ^{ξ}_{jk} is a covariance between two functional principal-component scores j and j'. T_{FPCA} follows a chi-square distribution with k degrees of freedom.

Other investigators [Zhu and Xiong 2012] proposed the PedCMC method, which is an extension of the CMC method for unrelated samples to pedigree samples. PedCMC statistic is defined as

$$T_{\mathrm{PedCMC}}=\frac{\frac{n_{case}(n-n_{case})}{n}\left[(\overline{V}_{\mathrm{case}}-\overline{V}_{\mathrm{cont}})^{\mathrm{T}}\Sigma_{k_1}^{-1}(\overline{V}_{\mathrm{case}}-\overline{V}_{\mathrm{cont}})+(\overline{G}_{\mathrm{case}}-\overline{G}_{\mathrm{cont}})^{\mathrm{T}}\Sigma_{k_2}^{-1}(\overline{G}_{\mathrm{case}}-\overline{G}_{\mathrm{cont}})\right]}{\left(D_r-\frac{n_{case}}{n}D_p\right)^{\mathrm{T}}\mathbf{\Phi}\left(D_r-\frac{n_{case}}{n}D_p\right)},$$

where $\overline{V}_{\mathrm{case}}$ and $\overline{V}_{\mathrm{cont}}$ are the average of the indicator variables in the case group and control group, respectively. $\overline{G}_{\mathrm{case}}$ and $\overline{G}_{\mathrm{cont}}$ are the average of the indicator variables for the genotype in case group and control group, respectively. The k variants are classified as k_1 groups of rare variants and k_2 common variant sites. Σ_{k_1} is denoted by $\Sigma_{k_1}=\mathrm{diag}\left(\sigma_1^2,\ \sigma_2^2,\ \ldots,\ \sigma_{k_1}^2\right)$, and Σ_{k_2} is defined as

$$\Sigma_{k_2}=\begin{bmatrix}\sigma_{11} & \cdots & \sigma_{1k_2}\\ \vdots & \ddots & \vdots\\ \sigma_{k_2 1} & \cdots & \sigma_{k_2k_2}\end{bmatrix}.$$

T_{PedCMC} follows a chi-square distribution with (k_1+k_2) degrees of freedom.

The RV-TDT methods [He, et al. 2014] are extensions of the TDT: they combine CMC, WSS, Burden of Rare Variants (BRV) [Auer, et al. 2013], and VT method. For parent i with variant j, we define c_{ij} and b_{ij} as

$$c_{ij}=\begin{cases}1 & \text{if a minor allele transmitted event occurs for parent } i \text{ with variant } j\\ 0 & \text{otherwise}\end{cases}$$

$$b_{ij}=\begin{cases}1 & \text{if a major allele transmitted event occurs for parent } i \text{ with variant } j\\ 0 & \text{otherwise}\end{cases}.$$

For genetic region k, all events are expressed as $c_i = \sum_{j=1}^{k} c_{ij}$ and $b_i = \sum_{j=1}^{k} b_{ij}$. Then, the TDT statistic is defined as

$$\chi^2=\frac{(b-c)^2}{b+c}.$$

For the TDT-CMC method, c and b are given by $c = \sum_{i=1}^{2n} \frac{c_i}{b_i+c_i}$ and $b = \sum_{i=1}^{2n} \frac{b_i}{b_i+c_i}$, where n is the number of trios. For the TDT-WSS method, c and b are expressed as $c = \sum_{i=1}^{2n} \sum_{j=1}^{k} c_{ij}/\hat{w}_j$ and $b = \sum_{i=1}^{2n} \sum_{j=1}^{k} b_{ij}/\hat{w}_j$, where weight is the estimated standard deviation of the number of variants in the parental haplotypes. For TDT-BRV, c and b are given by $c = \sum_{i=1}^{2n} c_i$ and $b = \sum_{i=1}^{2n} b_i$. The TDT-VT can be calculated by means of either the TDT-CMC or TDT-BRV.

Some investigators [Ma, et al. 2015] proposed burden test to analyze rare X-linked variant. The burden score statistic is

$$Q_{Burden} = \left(\sum_{j=1}^{k} w_j S_j\right)^2,$$

where $S_j = \sum_{i=1}^{n} G_{ij}\left(y_i - \hat{\mu}_i\right)$ is the score statistic, and $\hat{\mu}_i$ is the estimated mean of y_i under the null hypothesis. For females, let G_{ij} = (0, 1, 2) be the vector of genotypes. For males, two coding schemes are considered: G_{ij} = (0, 2) when assuming XCI and G_{ij} = (0, 1) when assuming E-XCI.

In a family-based study, PedCMC-Burden for the X chromosome have been proposed [Schaid, et al. 2013]. The statistic for the type of burden test is

$$\mathrm{T} = \frac{\left[\left(\mathrm{Y} - \hat{\mathrm{Y}}\right)'S\right]^2}{\left(\mathrm{Y} - \hat{\mathrm{Y}}\right)'V_S\left(\mathrm{Y} - \hat{\mathrm{Y}}\right)}.$$

The matrix V_S has elements $\mathrm{Cov}(S_i, S_j) = \alpha_{ij}\Omega_{ij}c_s$ where

$$\alpha_{ij} = \begin{cases} 2 & \text{if female-female} \\ d^2 & \text{if male-male} \\ d & \text{if female-male} \end{cases},$$

and

$$c_S = \sum_{k_1=1}^{k}\sum_{k_2=1}^{k} w_{k_1}w_{k_2}R_{k_1k_2}\sqrt{p_{k_1}\left(1 - p_{k_1}\right)p_{k_2}\left(1 - p_{k_2}\right)}.$$

Let $R_{k_1k_2}$ denote the correlation of genetic scores between k_1 and k_2, and Ω_{ij} is the X chromosome kinship coefficients for subjects i and j.

2.2.2 Variance component-type methods

Variance component-type approaches within a random effect model such as a C-alpha test [Neale, et al. 2011] and SKAT [Wu, et al. 2011] analysis for an association by evaluating the distribution of genetic effects in a group of rare variants. For the family-based design, family-based SKAT (famSKAT) has been proposed [Chen, et al. 2013]. For rare X-linked variant association analysis, extensions of SKAT [Ma, et al. 2015] and PedCMC-Kernel [Schaid, et al. 2013] have been proposed in case-control design and family-based design, respectively.

The C-alpha [Neale, et al. 2011] statistic is

$$T_{\text{C-alpha}} = \sum_{j=1}^{k} \left\{ \left(n_j^{case} - n_j p_0 \right)^2 - n_j p_0 \left(1 - p_0 \right) \right\},$$

where n_j^{case} is the number of rare variants at variant j in cases, and n_j is the number of rare variants at j-th variants; p_0 is denoted by $n_{case}/(n_{case} + n_{cont})$. The p-value of C-alpha is calculated using a permutation procedure. The C-alpha method is robust in the presence of deleterious and protective variants.

The C-alpha test is a special case of the SKAT method [Wu, et al. 2011]. The SKAT is a weighted sum of single variant score statistics. The SKAT statistic is

$$Q_{SKAT} = (\text{Y} - \hat{\mu})'\text{GWG}'(\text{Y} - \hat{\mu}) = \sum_{j=1}^{k} w_j^2 S_j^2 ,$$

where the weight, w_j, is assumed to be a function of MAF via the beta density function Beta(MAF_j, 1, 25). S_j is a single variant score statistic:

$$S_j = \sum_{i=1}^{n} G_{ij}(y_i - \hat{\mu}_i),$$

where $\hat{\mu}_i$ is the estimated mean of y_i under the null hypothesis. Q_{SKAT} asymptotically follows a mixture of chi-square distribution under the null hypothesis. The *p*-values can be calculated by the Davies method [Davies 1980a].

Some researchers [Chen, et al. 2013] proposed the FamSKAT method, which is an extension of the SKAT method for unrelated samples to pedigree samples. The famSKAT test statistic is

$$Q_{famSKAT} = (\mathrm{Y} - \hat{\mu})'\hat{\Sigma}^{-1}\mathrm{GWG}'\hat{\Sigma}^{-1}(\mathrm{Y} - \hat{\mu}),$$

where Σ is denoted by $h^2\Phi+(1- h^2)\mathrm{I}$. h^2 is the heritability of the trait. The *p*-value of $Q_{famSKAT}$ is calculated using a moment matching method [Liu, et al. 2009].

Another research group [Ma, et al. 2015] proposed SKAT to analyze rare X-linked variant. The SKAT score statistic is

$$Q_{SKAT} = \sum_{j=1}^{k} w_j^2 S_j^2,$$

and follows a mixture of chi-square distributions.

In a family-based study, PedCMC-Kernel for the X chromosome have been proposed [Schaid, et al. 2013]. The quadratic kernel statistic can be expressed as

$$\mathrm{Q} = \sum_{j=1}^{k} \left[w_j \sum_{i=1}^{n} (y_i - \hat{y}_i) G_{ij} \right]^2 .$$

The distribution of Q follows as scaled chi-square distribution, with scale and degrees of freedom estimated by the first two moment. The scale is estimated as $\delta = Var(\mathrm{Q})/[2E(\mathrm{Q})]$, and the degree of freedom as $d = 2E(\mathrm{Q})^2/Var(\mathrm{Q})$. The p-value of PedCMC-Kernel is computed by assuming $Q_{scaled} = \mathrm{Q}/\delta \sim \chi^2_d$.

2.2.3 Optimal-type methods

Optimal-type approaches such as SKAT-O [Lee, et al. 2012a], Fisher's hybrid statistics [Derkach, et al. 2013], and Mixed effects Score Test (MiST) [Sun, et al. 2013] have been proposed recently. For the family-based design, the minimum *p*-value optimized nuisance parameter score test extended to relatives (MONSTER) was proposed [Jiang and McPeek 2014]. For rare X-linked variant association analysis, an extension of SKAT-O [Ma, et al. 2015] has been proposed in case-control design.

The SKAT-O method [Lee, et al. 2012a] is a linear combination of SKAT and burden test statistics. The SKAT-O test statistic takes the form

$$Q_\rho = (1 - \rho)Q_{SKAT} + \rho Q_{burden}, \quad 0 \le \rho \le 1.$$

The SKAT-O test is reduced to the SKAT when $\rho = 0$ and to the burden test when $\rho = 1$. The *p*-value assumes the smallest *p*-value across the values of ρ and can be obtained with the mixture of chi-square distribution with a single degree of freedom.

Another approach is to use Fisher's method of combining *p*-values from SKAT and burden statistics [Derkach, et al. 2013]. The Fisher statistic is defined as

$$T_{Fisher} = -2\log(p_{SKAT}) - 2\log(p_{burden}).$$

The *p*-value of T_{Fisher} follows a chi-square distribution with four degrees of freedom.

Some investigators [Sun, et al. 2013] proposed the MiST method, which is modified SKAT test statistic to make it independent from the burden test statistic and derived the asymptotic *p*-value of the Fisher method.

Another research group [Jiang and McPeek 2014] proposed the MONSTER method, which is an extension of SKAT-O method for unrelated samples to pedigree samples. The MONSTER test statistic is a convex combination of famSKAT and famBT [Chen, et al. 2013]. It can be written as

$$T_\rho = (1 - \rho)Q_{famSKAT} + \rho Q_{famBT}, \quad 0 \le \rho \le 1.$$

Including the ρ parameter allows MONSTER to balance between two statistics in order to achieve robustness for a wide range of possible genetic architectures of the trait.

Another investigators [Ma, et al. 2015] proposed SKAT-O to analyze rare X-linked variant. The SKAT-O statistic is a weighted average of Q_{Burden} and Q_{SKAT}. We perform a grid search to estimate the optimal ρ, and select ρ parameter such that the Q_ρ is maximized.

2.2.4 Other methods

Many other methods such as Replication based test (RBT) [Ionita-Laza, et al. 2011], Variant Annotation, Analysis & Search Tool (VAAST) [Yandell, et al. 2011], Exponential-Combination (EC) test [Chen, et al. 2012], and Optimal Combination of Single-variant Test (OCST) [Sha and Zhang 2014] have been proposed. For the family-based design, pedigree-VAAST (pVAAST) [Hu, et al. 2014] has been proposed recently.

The RBT [Ionita-Laza, et al. 2011] statistic can be written as:

$$S = \sum_{k=0}^{N_r} \sum_{k' > k} - n_k^{k'} \log[p(k, k')] ,$$

where Nr is an upper threshold on the number of occurrences of a variant in controls, and $n_k^{k'}$ is the size of group (k, k'). The RBT method is based on a weighted-sum statistic, but that has advantage of being less sensitive to the presence of both risk and protective variants.

Some investigators [Ionita-Laza, et al. 2011] proposed VAAST method to achieve robust power in the presence of both protective and harmful variants. The VAAST combines variant frequency data with Amino Acid Substitution (AAS) effect information on a feature-by-features basis using the likelihood ratio. The likelihood ratio, λ, is equal to:

$$\lambda = \ln\left(\frac{L_{Null}}{L_{Alt}}\right) = \sum_{j=1}^{k+m} \ln\left[\frac{h_j \left(\hat{p}_j\right)^{X_j} \left(1 - \hat{p}_j\right)^{2l_j n_j - X_j}}{a_j \left(\hat{p}_j^U\right)^{X_j^U} \left(1 - \hat{p}_j^U\right)^{2l_j n_j^U - X_j^U} \left(\hat{p}_j^A\right)^{X_j^A} \left(1 - \hat{p}_j^A\right)^{2l_j n_j^A - X_j^A}}\right],$$

where h_j is the proportion of the type of amino acid change in the population background, and a_j is the proportion of the type of amino acid change among all disease-causing mutations in OMIM.

The EC test [Chen, et al. 2012] test statistic is

$$Q_{EC} = \sum_{j=1}^{k} \pi_j \exp\left(w_j Z_j^2\right),$$

where π_j is the weight on the linear scale for the individual variant statistic Z_j, and w_j is the weight on exponential scale. Since the null distribution of Q_{EC} is unknown, the *p*-value is calculated by a permutation method.

Some researchers proposed [Sha and Zhang 2014] an OCST by combining information from the tests of the three classes. The OCST is defined as

$$T_{OCST} = \min_{z \in [1, \infty]} \left(P_a(z), P_b(z), P_c(z)\right),$$

where $P_a(z)$, $P_b(z)$, and $P_c(z)$ denote the *p*-values of $T_a(z)$, $T_b(z)$, and $T_c(z)$, respectively. The three test classes are 'only risk variants', 'both risk and protective variants', and 'only protective variants'.

In a family-based study, pVAAST [Hu, et al. 2014] combines the existing variant prioritization and case-control association features in VAAST with a new linkage analysis method. pVAAST support a variety of simulated and real data sets involving dominant, recessive, and *de novo* inheritance models. pVAAST CLRT ($CLRT_p$) score can be written as:

$$\mathrm{CLRT_p} = c \sum_{i=1}^{n} LOD_i - 2\lambda ,$$

where LOD_i is the lod score for the i-th family and $c = 2 \times \ln(10)$. The statistical significance is calculated by using a combination of permutation and gene-drop simulation to account for both the family structure and the observed pattern of variation in cases and controls.

This chapter was published in *Bioinformatics*
as a partial fulfillment of Sungkyoung Choi's Ph.D program.

Chapter 3

Family-based rare variant association test

3.1 Introduction

In this chapter, we propose a *FA*mily-based *R*are *V*ariant *A*ssociation *T*est (***FARVAT***). We provide a burden test and a variance component test for extended families, and these approaches are extended to the SKAT-O-type statistic. The proposed method assumes that families are ascertained based on the disease status of family members, and minor allele frequencies (MAFs) between affected and unaffected individuals are compared. MAFs for each rare variant are estimated with the best linear unbiased estimators [McPeek, et al. 2004]. *FARVAT* is implemented with C++ and is computationally efficient for the analysis of rare variants with extended families. With extensive

simulations, we compared the proposed methods with existing methods [He, et al. 2014; Zhu and Xiong 2012], and results showed that the proposed methods were the most efficient in the considered scenarios. Application of the proposed method to schizophrenia and GAW17 illustrated its practical value in real analyses.

3.2 Methods

3.2.1 Notations and the disease model

We assumed that there are n families and n_i individuals in family i, and the total sample size was denoted by $N = \sum_{i=1}^{n} n_i$. We assumed that genotype data for m rare variant loci were available. We let y_{ij} and x_{ij}^{k} be the phenotype and genotype count of an individual j in a family i for rare variant k. If we denoted the disease prevalence by q, y_{ij} was coded as 1 for affected individuals, q for individuals with missing phenotype, and 0 for unaffected individuals. If genotype frequencies of affected and unaffected individuals are compared to detect genetic associations, the statistical efficiency can be improved by modifying the phenotype [Lange and Laird 2002; Thornton and McPeek 2007], and we therefore introduced the so-called offset μ_{ij} to set $t_{ij} = y_{ij} - \mu_{ij}$. The disease prevalence q has often been used as an offset, and if the disease prevalences in males and females are different, the offset should be chosen separately [Thornton, et al. 2012]. For randomly selected families, the best linear unbiased predictor (BLUP) from the linear mixed model is

known to be an efficient choice for μ_{ij} [Won and Lange 2013]. With this choice of offset, the effects of covariates can properly be adjusted. Then, if we set the column vectors that comprise x_{ij}^{k} and t_{ij} for individuals in a family i by $\boldsymbol{X}_i^k$ and $\boldsymbol{T}_i$ respectively, we denoted

$$\mathbf{X}^k = \begin{pmatrix} \mathbf{X}_1^k \\ \vdots \\ \mathbf{X}_n^k \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} \mathbf{X}^1 & \cdots & \mathbf{X}^m \end{pmatrix}, \text{ and } \mathbf{T} = \begin{pmatrix} \mathbf{T}_1 \\ \vdots \\ \mathbf{T}_n \end{pmatrix}.$$

The variance-covariance matrix of $\mathbf{X}^k$ for extended families could be calculated based on the kinship coefficient. If we let $\phi_{ij,i'j'}$ be the kinship coefficient between individuals j in a family i and j' in a family i', and let d_{ij} be the inbreeding coefficient for an individual j in family i, $\boldsymbol{\Phi_i}$ was denoted by

$$\begin{pmatrix} 1+d_{i1} & 2\phi_{i1,i2} & \cdots & 2\phi_{i1,in_i} \\ 2\phi_{i2,i1} & 1+d_{i2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 2\phi_{i(n_i-1),in_i} \\ 2\phi_{in_i,i1} & \cdots & 2\phi_{in_i,i(n_i-1)} & 1+d_{in_i} \end{pmatrix},$$

and we let

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\Phi_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\Phi_n} \end{pmatrix}.$$

If we denote the covariance between x_{ij}^{k} and $x_{ij}^{k'}$ by $\sigma_{kk'}$, we have cov($\mathbf{X}^k$, $\mathbf{X}^{k'}$) = $\sigma_{kk'}\boldsymbol{\Phi}$, and $\sigma_{kk'}$ is estimated with the empirical covariance.

In the presence of population substructure, $\boldsymbol{\Phi}$ should be empirically estimated with common variants available at the genome-wide level instead of

using the kinship coefficient between individuals [Thornton and McPeek 2010]. We assume that there are A common variants and the coded genotype for common variant is denoted by x'^{a}_{ij} for individual j in family i at common variant a. If we let p_a be the minor allele frequency of common variant a, $\phi_{ij,i'j'}$ for $\mathbf{\Phi}$ [Thornton and McPeek 2010] can be estimated by

$$\phi_{ij,i'j'} = \begin{cases} \frac{1}{A}\sum_{a=1}^{A} \frac{(x'^{a}_{ij} - 2p_a)(x'^{a}_{i'j'} - 2p_a)}{2p_a(1-p_a)}, & i \neq i' \text{ and } j \neq j' \\ 1 + \frac{1}{A}\sum_{a=1}^{A} \frac{x'^{a}_{ij} - (1+2p_a)x'^{a}_{ij} + 2p_a^2}{2p_a(1-p_a)}, & \text{ow.} \end{cases}$$

3.2.2 Family-based rare variant association test

For ascertained samples, the disease status can be assumed to be fixed, and the genotype frequencies between affected and unaffected individuals are usually compared. We let $\mathbf{1}_w$ be the $w\times1$ column vector that consisted of 1 and $\mathbf{I}_w$ be the $w\times w$ identity matrix. If we denoted an MAF of rare variant k in unaffected individuals by p_k, we assumed [Thornton and McPeek 2007] that for a constant γ_k,

$$E(\mathbf{X}^k \mid \mathbf{Y}) = 2p_k\mathbf{1}_N + \gamma_k\mathbf{Y},\ \text{var}(\mathbf{X}^k \mid \mathbf{Y}) = \sigma_{kk}\mathbf{\Phi},$$

where $0 < 2p_k + \gamma_k < 1$. If we let $\mathbf{V}$ be the working variance-covariance matrix, the score for the quasi-likelihood [Thornton and McPeek 2007] became

$$\mathbf{T}'\mathbf{V}^{-1}(\mathbf{X} - E(\mathbf{X})).$$

Recently we showed that the approximate optimal efficiency for the analysis of common variants is achieved with $\mathbf{V} = \mathbf{I}_N$ [Won and Lange 2013]. For the choice of the offset in $\mathbf{T}$, BLUP and q have been suggested for randomly selected samples and ascertained samples, respectively [Thornton and McPeek 2007; Won and Elston 2008]. $E(\mathbf{X})$ can be estimated with the following best linear unbiased estimator [McPeek, et al. 2004]:

$$\hat{E}(\mathbf{X}) = \mathbf{1}_N(\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\mathbf{X}.$$

Therefore, our score based on the quasi-likelihood became

$$\mathbf{T}^t(\mathbf{I}_N - \mathbf{1}_N(\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\boldsymbol{\Phi}^{-1})\mathbf{X}.$$

If we let

$$\mathbf{H} = \boldsymbol{\Phi} - \mathbf{1}_N(\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & \ddots & \vdots \\ \sigma_{m1} & \cdots & \sigma_{mm} \end{pmatrix},$$

we have

$$\text{var}\left(\mathbf{T}^t\left(\mathbf{I}_N - \mathbf{1}_N(\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\right)\mathbf{X}^k\right) = \sigma_{kk}\mathbf{T}^t\mathbf{H}\mathbf{T},$$

and thus the variance-covariance matrix of the score was

$$\text{var}\left(\mathbf{T}^t\left(\mathbf{X}^1 - \hat{E}(\mathbf{X}^1)\right) \quad \cdots \quad \mathbf{T}^t\left(\mathbf{X}^m - \hat{E}(\mathbf{X}^m)\right)\right) = \left(\mathbf{T}^t\mathbf{H}\mathbf{T}\right)\boldsymbol{\Sigma}.$$

Therefore, we have

$$\frac{1}{\sqrt{\mathbf{T}^t\mathbf{H}\mathbf{T}}}\mathbf{T}^t(\mathbf{I}_N - \mathbf{1}_N(\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\boldsymbol{\Phi}^{-1})\mathbf{X}\boldsymbol{\Sigma}^{-1/2} \sim MVN(\mathbf{0}, \mathbf{I}_m) \text{ under } H_0.$$

For rare variant association analysis, the collapsed amount of either rare alleles or variance inflation between affected and unaffected individuals has been compared [Li and Leal 2008; Neale, et al. 2011; Price, et al. 2010; Wu, et al. 2011]. If we let the weight for variant k be w_k, the null hypothesis for the former was

$$H_0^1 : w_1\gamma_1 + ... + w_m\gamma_m = 0,$$

and that for the latter was

$$H_0^2 : w_1^2\gamma_1^2 + ... + w_m^2\gamma_m^2 = 0.$$

For the choice of w_k, $w_k = 1$ or $[p_k(1-p_k)]^{-1/2}$ are often utilized. If we denoted the $m \times m$ diagonal matrix, which consists of w_k, by $\mathbf{W}$, the score test for the burden-type test was

$$\frac{1}{\mathbf{T}^t\mathbf{H}\mathbf{T}}\mathbf{T}^t(\mathbf{X}-\hat{E}(\mathbf{X}))\mathbf{W}\mathbf{1}_m\mathbf{1}_m^t\mathbf{W}(\mathbf{X}-\hat{E}(\mathbf{X}))^t\mathbf{T},$$

and the score test for the C-alpha-type test was

$$\frac{1}{\mathbf{T}^t\mathbf{H}\mathbf{T}}\mathbf{T}^t(\mathbf{X}-\hat{E}(\mathbf{X}))\mathbf{W}\mathbf{I}_m\mathbf{W}(\mathbf{X}-\hat{E}(\mathbf{X}))^t\mathbf{T}.$$

Both score tests for rare variant analysis could be generalized to

$$\frac{1}{\mathbf{T}^t\mathbf{H}\mathbf{T}}\mathbf{T}^t(\mathbf{X}-\hat{E}(\mathbf{X}))\mathbf{W}\mathbf{R}\mathbf{W}(\mathbf{X}-\hat{E}(\mathbf{X}))^t\mathbf{T},$$

and for a given constant $c \in [0,1]$, S_c was denoted by

$$\frac{1}{\mathbf{T}^t\mathbf{H}\mathbf{T}}\mathbf{T}^t(\mathbf{X}-\hat{E}(\mathbf{X}))\mathbf{W}\left((1-c)\mathbf{I}_m + c\mathbf{1}_m\mathbf{1}_m^t\right)\mathbf{W}(\mathbf{X}-\hat{E}(\mathbf{X}))^t\mathbf{T}.$$

We denoted eigenvalues for $\mathbf{\Sigma}^{1/2}\mathbf{W}\mathbf{W}\mathbf{\Sigma}^{1/2}$ by λ_k. If we let χ^2_k's be independent chi-square distributions with a single degree of freedom, we have

$$S_1 \sim (\mathbf{1}_m^t \mathbf{W}\mathbf{\Sigma}\mathbf{W}\mathbf{1}_m)\chi_1^2 \text{ under } H_0^1,$$

and

$$S_0 \sim \sum_{k=1}^{m} \lambda_k \chi_k^2 \text{ under } H_0^2.$$

The p-values for S_1 and S_0 were respectively denoted by ***FARVAT$_b$*** and ***FARVAT$_c$***, and in particular, ***FARVAT$_c$*** can be calculated with the Davies method [Davies 1980b] or the method described by Liu et al [Lee, et al. 2012b; Liu, et al. 2009].

3.2.3 Extension to the optimal type statistic

The burden test is known to be efficient if all rare variants have either deleterious or protective effects on disease; otherwise, the C-alpha test is more efficient [Neale, et al. 2011]. A balanced approach for both scenarios can be achieved by the SKAT-O type statistic [Lee, et al. 2012b]. For $c_0 = 0 < c_1 < \ldots < c_L = 1$, we denoted the observed value for S_{c_l} by s_{c_l}, and their corresponding p-values were denoted by p_{c_l}. Furthermore, we denoted the $(1-p)th$ quantile for S_{c_l} by $Q_{c_l}(p)$. If we let

$$p_{\min} = \min\left\{p_{c_0}, p_{c_1}, \ldots, p_{c_L}\right\},$$

our final p-value was obtained by

$$1-P(S_{c_0}\le Q_{c_0}(p_{\min}),...,S_{c_L}\le Q_{c_L}(p_{\min})).$$

The numerical calculation of the final *p*-value for the independent samples was derived by Lee et al [Lee, et al. 2012a; Lee, et al. 2012b], and our final *p*-values, denoted by $FARVAT_o$ were calculated based on their approach with some modification.

If we let $\mathbf{Z}=\mathbf{\Sigma}^{1/2}\mathbf{W}$ and $\bar{\mathbf{Z}}=\mathbf{Z}1_m(1_m^t1_m)^{-1}$, the projection matrix onto a space spanned by $\bar{\mathbf{Z}}$ becomes $\mathbf{\Pi}=\bar{\mathbf{Z}}(\bar{\mathbf{Z}}^t\bar{\mathbf{Z}})^{-1}\bar{\mathbf{Z}}^t$. If we let

$$\mathbf{u}=\frac{1}{\sqrt{\mathbf{T}^t\mathbf{H}\mathbf{T}}}\mathbf{T}^t(\mathbf{I}_N-\mathbf{1}_N(\mathbf{1}_N^t\mathbf{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\mathbf{\Phi}^{-1})\mathbf{X}\Sigma^{-1/2},$$

$\mathbf{u}\sim MVN(\mathbf{0}, \mathbf{I}_m)$ and S_{c_l} becomes

$$S_{c_l}=\mathbf{u}^t\Sigma^{1/2}\mathbf{W}\mathbf{R}\mathbf{W}\Sigma^{1/2}\mathbf{u}=\mathbf{u}^t\mathbf{Z}\mathbf{R}\mathbf{Z}^t\mathbf{u}=(1-c_l)\mathbf{u}^t\mathbf{Z}\mathbf{Z}^t\mathbf{u}+c_lm^2\mathbf{u}^t\bar{\mathbf{Z}}\bar{\mathbf{Z}}^t\mathbf{u}.$$

As was shown by Lee et al[Lee, et al. 2012a; Lee, et al. 2012b], if we let

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

we have

$$\begin{aligned}S_{c_l}&=(1-c_l)\mathbf{u}^t(\mathbf{I}_m-\mathbf{\Pi})\mathbf{Z}\mathbf{Z}^t(\mathbf{I}_m-\mathbf{\Pi})\mathbf{u}\\&\quad+2(1-c_l)\mathbf{u}^t(\mathbf{I}_m-\mathbf{\Pi})\mathbf{Z}\mathbf{Z}^t\mathbf{\Pi}\mathbf{u}+\tau(c_l)\mathbf{u}^t\mathbf{\Pi}\mathbf{u},\end{aligned}$$

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

$$\begin{aligned}&P(S_{c_0}\le Q_{c_0}(p_{\min}),...,S_{c_L}\le Q_{c_L}(p_{\min}))\\&=E\left\{P\left(S_{c_0}\le Q_{c_0}(p_{\min}),...,S_{c_L}\le Q_{c_L}(p_{\min}))\,|\,\mathbf{u}^t\mathbf{\Pi}\mathbf{u}=\eta\right)\right\},\end{aligned}$$

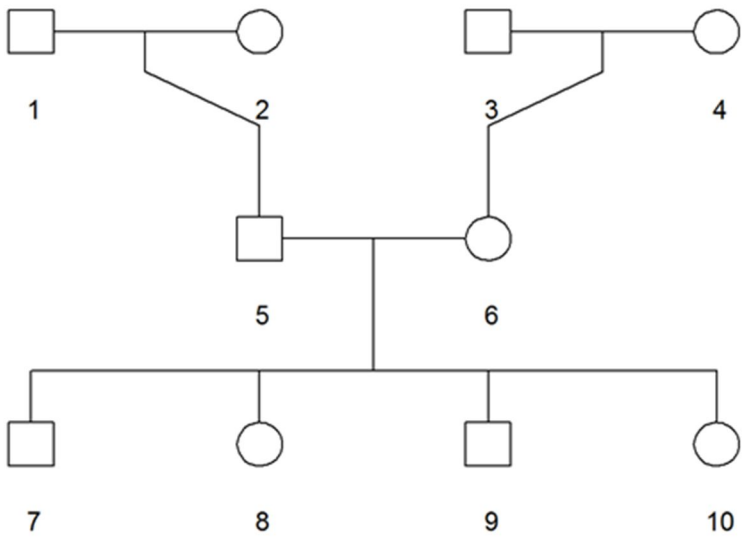
and the following conditional probability can be numerically calculated, as was suggested by Lee et al[Lee, et al. 2012a; Lee, et al. 2012b]:

3.3 Simulation study

3.3.1 The simulation model

In our simulation studies, we considered extended families that consisted of 10 individuals, and extended over three generations (see Figure 3.1). To generate the genotypes for extended families, haplotypes were simulated with COSI software [Schaffner, et al. 2005], based on the coalescent model, and obtained haplotypes were used for founders' genotypes. In the coalescent model for COSI, we assumed that the mutation rate was 1.5×10^{-8}, and 5,000 haplotypes with 50,000 base pairs were generated. *m* rare variants in a region or all rare variants for which MAFs were less than 0.01 were randomly selected, and pairs of haplotypes were randomly chosen with replacement to derive the founders' genotypes. Under the assumption of no recombination, a haplotype from each founder was randomly selected to construct non-founders' genotypes under the assumption of Mendelian transmission. The disease status for each individual was generated with the liability threshold model. The underlying liabilities were defined by summing the phenotypic mean, polygenic effect, common environmental effect, main genetic effect, and random error.

Figure 3.1 Extended family used in our simulation studies



The phenotypic mean β_0 was assumed to be 0, and the polygenic effect, common environmental effect, and random errors were generated from the normal distribution with mean 0. Variances for the polygenic effect, common environmental effect, and random errors were denoted by σ_g^2, σ_c^2, and σ_e^2, respectively, and were assumed to be 1. In this setting, the heritability was 1/3. The polygenic effect was independently generated from $N(0, \sigma_g^2)$ for founders, and the average of maternal and paternal polygenic effects was combined with values independently sampled from $N(0, 0.5\sigma_g^2)$ for the polygenic effects of offspring. Common environmental effects were assumed to be the same for all individuals in each family. We assumed there were m rare variants, and their main genetic effects for each individual were the product of β_k and the number of disease alleles. If we let h_a^2 be the relative proportion of variance explained by rare variants, β_k were sampled from $U(1.0, v)$ and v was calculated by

$$v = \sqrt{\frac{(\sigma_g^2 + \sigma_c^2 + \sigma_e^2)h_a^2}{(1-h_a^2)\sum_{k=1}^{m}\beta_k^2 2p_k(1-p_k)}}.$$

Under the null hypothesis, h_a^2 was set to 0, and β_k became 0. Once the underlying liabilities of main genetic effects, polygenic effects, common environmental effects, and random errors were generated, they were transformed to being affected if they were larger than the threshold; otherwise, they were considered as unaffected. The threshold was chosen to preserve the assumed prevalence, and the disease prevalence was assumed to be 0.12. Families with more than two affected grandchildren were utilized for

simulation studies, and sampling was repeated until the given numbers of these families were obtained.

Furthermore, the robustness of the proposed statistic under the presence of the population substructure was evaluated with simulated data. We assumed that there were two subpopulations, and each founder was assigned to the one of the two subpopulations with 50% probability. Means of liabilities for phenotypes in both populations differed by 0.2. The allele frequencies for each marker in the two subpopulations were generated by the Balding-Nichols model [Balding and Nichols 1995]. That is, for marker k, the allele frequency, p_k, in an ancestral population was generated from $U(0.0001, 0.01)$, and the marker allele frequencies for the two subpopulations were independently sampled from the beta distributions $(p_k(1 - F_{ST})/F_{ST}, (1- p_k)(1 - F_{ST})/F_{ST})$. A survey reported F_{ST} estimates with a median of 0.008 and a 90th percentile of 0.028 among Europeans; the corresponding values were 0.027 and 0.14, respectively, among Africans, and 0.043 and 0.12, respectively, among Asian [Cavalli-Sforza and Piazza 1993]. The values for Wright's F_{ST} were assumed to be 0.005, 0.01, and 0.05.

3.3.2 Evaluation with simulated data under the absence of population substructure

The statistical validity of $FARVAT_b$, $FARVAT_c$, and $FARVAT_o$ was evaluated under the absence of population substructure, and the results were compared with PedCMC, FPCA [Zhu and Xiong 2012], and RV-TDT methods [He, et al. 2014]. RV-TDT methods consist of BRV.T01, BRV.Hapo.T01, CMC.T01, CMC.Hapo.T01, VT.BRV.Hapo, VT.CMC.Hapo and WSS.Hapo. We generated 50 and 100 extended families in each replicate, and empirical type-1 error estimates at the 0.05 and 0.01 significance levels were calculated with 50,000 replicates. For the proposed methods, 1 and $[p_k(1 - p_k)]^{-1/2}$ were considered for w_k, and the kinship coefficients were used to build the correlation matrix $\mathbf{\Phi}$. Rare variants for which MAFs were less than 0.01 were considered for all statistics. Tables 3.1 and 3.2, 30 and 100 rare variants were randomly selected, and in Table 3.3, all rare variants in the 30 kb genetic region were considered. These results showed that the empirical type-1 error estimates for $FARVAT_b$, $FARVAT_c$, and $FARVAT_o$ preserved the nominal significance levels. However, CMC.T01, BRV.Hapo.T01, CMC.Hapo.T01, VT.BRV.Hapo, VT.CMC.Hapo, WSS.Hapo, and FPCA were usually conservative, and BRV.T01 and PedCMC seemed to be liberal. For PedCMC, genotype scores of individuals with more than or equal to a single rare allele were considered as 1; otherwise they were 0.

Table 3.1 Empirical type-1 error estimates for 30 rare variants. The empirical type-1 error estimates were calculated with 50,000 replicates at the 0.05 and 0.01 significance levels. MAFs for all variants are assumed to be less than 0.01, and 30 rare variants are randomly selected. The numbers of families are denoted by *n*.

n	statistics	weight	α=.05(95%CI)	α=.01(95%CI)
50	BRV.T01	1	0.050 ± 0.002	0.011 ± 0.001
	CMC.T01	1	0.038 ± 0.002	0.006 ± 0.001
	BRV.Hapo.T01	1	0.033 ± 0.002	0.005 ± 0.001
	CMC.Hapo.T01	1	0.033 ± 0.002	0.005 ± 0.001
	VT.BRV.Hapo	1	0.036 ± 0.002	0.006 ± 0.001
	VT.CMC.Hapo	1	0.034 ± 0.002	0.006 ± 0.001
	WSS.Hapo	1	0.022 ± 0.001	0.004 ± 0.001
	PedCMC	1	0.051 ± 0.002	0.009 ± 0.001
	FPCA	1	0.033 ± 0.002	0.006 ± 0.001
	$FARVAT_b$	1	0.055 ± 0.002	0.011 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.055 ± 0.002	0.010 ± 0.001
	$FARVAT_c$	1	0.037 ± 0.002	0.005 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.041 ± 0.002	0.006 ± 0.001
	$FARVAT_o$	1	0.043 ± 0.002	0.007 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.043 ± 0.002	0.008 ± 0.001
100	BRV.T01	1	0.049 ± 0.002	0.011 ± 0.001
	CMC.T01	1	0.037 ± 0.002	0.006 ± 0.001
	BRV.Hapo.T01	1	0.030 ± 0.001	0.005 ± 0.001
	CMC.Hapo.T01	1	0.030 ± 0.001	0.005 ± 0.001
	VT.BRV.Hapo	1	0.034 ± 0.002	0.006 ± 0.001
	VT.CMC.Hapo	1	0.033 ± 0.002	0.005 ± 0.001
	WSS.Hapo	1	0.017 ± 0.001	0.003 ± 0.000
	PedCMC	1	0.050 ± 0.002	0.010 ± 0.001
	FPCA	1	0.030 ± 0.001	0.006 ± 0.001
	$FARVAT_b$	1	0.051 ± 0.002	0.010 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.051 ± 0.002	0.010 ± 0.001
	$FARVAT_c$	1	0.039 ± 0.002	0.006 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.044 ± 0.002	0.007 ± 0.001
	$FARVAT_o$	1	0.043 ± 0.002	0.008 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.044 ± 0.002	0.008 ± 0.001

Table 3.2 Empirical type-1 error estimates for 100 rare variants. The empirical type-1 error estimates were calculated with 50,000 replicates at the 0.05 and 0.01 significance levels. MAFs for all variants are assumed to be less than 0.01, and 100 rare variants are randomly selected. The numbers of families are denoted by *n*.

n	statistics	weight	α=.05(95%CI)	α=.01(95%CI)
50	BRV.T01	1	0.060 ± 0.002	0.017 ± 0.0011
	CMC.T01	1	0.020 ± 0.001	0.003 ± 0.0005
	BRV.Hapo.T01	1	0.014 ± 0.001	0.002 ± 0.0004
	CMC.Hapo.T01	1	0.010 ± 0.001	0.001 ± 0.0003
	VT.BRV.Hapo	1	0.019 ± 0.001	0.003 ± 0.0005
	VT.CMC.Hapo	1	0.012 ± 0.001	0.002 ± 0.0004
	WSS.Hapo	1	0.007 ± 0.001	0.001 ± 0.0003
	PedCMC	1	0.052 ± 0.002	0.010 ± 0.001
	FPCA	1	0.028 ± 0.001	0.005 ± 0.001
	$FARVAT_b$	1	0.053 ± 0.002	0.011 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.054 ± 0.002	0.011 ± 0.001
	$FARVAT_c$	1	0.032 ± 0.002	0.004 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.037 ± 0.002	0.006 ± 0.001
	$FARVAT_o$	1	0.042 ± 0.002	0.007 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.040 ± 0.002	0.008 ± 0.001
100	BRV.T01	1	0.051 ± 0.002	0.015 ± 0.0011
	CMC.T01	1	0.016 ± 0.001	0.002 ± 0.0004
	BRV.Hapo.T01	1	0.007 ± 0.001	0.001 ± 0.0003
	CMC.Hapo.T01	1	0.005 ± 0.001	0.001 ± 0.0002
	VT.BRV.Hapo	1	0.014 ± 0.001	0.002 ± 0.0004
	VT.CMC.Hapo	1	0.009 ± 0.001	0.002 ± 0.0003
	WSS.Hapo	1	0.003 ± 0.000	0.000 ± 0.0002
	PedCMC	1	0.049 ± 0.002	0.010 ± 0.001
	FPCA	1	0.027 ± 0.001	0.005 ± 0.001
	$FARVAT_b$	1	0.053 ± 0.002	0.011 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.051 ± 0.002	0.011 ± 0.001
	$FARVAT_c$	1	0.036 ± 0.002	0.005 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.043 ± 0.002	0.007 ± 0.001
	$FARVAT_o$	1	0.044 ± 0.002	0.008 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.043 ± 0.002	0.009 ± 0.001

Table 3.3 Empirical type-1 error estimates in the 30 kb genetic region. The empirical type-1 error estimates were calculated with 50,000 replicates at the 0.05 and 0.01 significance levels. All rare variants of which MAFs are less than 0.01 are used to calculate each statistic. The numbers of families are denoted by *n*.

n	statistics	weight	α=.05(95%CI)	α=.01(95%CI)
50	BRV.T01	1	0.102 ± 0.003	0.031±0.002
	CMC.T01	1	0.071 ± 0.002	0.014 ± 0.001
	BRV.Hapo.T01	1	0.045 ± 0.002	0.009 ± 0.001
	CMC.Hapo.T01	1	0.043 ± 0.002	0.008 ± 0.001
	VT.BRV.Hapo	1	0.048 ± 0.002	0.009 ± 0.001
	VT.CMC.Hapo	1	0.045 ± 0.002	0.009 ± 0.001
	WSS.Hapo	1	0.025 ± 0.001	0.004 ± 0.001
	PedCMC	1	0.088 ± 0.002	0.021 ± 0.001
	FPCA	1	0.055 ± 0.002	0.012 ± 0.001
	$FARVAT_b$	1	0.052 ± 0.002	0.011 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.052 ± 0.002	0.010 ± 0.001
	$FARVAT_c$	1	0.046 ± 0.002	0.009 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.049 ± 0.002	0.011 ± 0.001
	$FARVAT_o$	1	0.056 ± 0.002	0.012 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.050 ± 0.002	0.011 ± 0.001
100	BRV.T01	1	0.090 ± 0.003	0.027 ± 0.0014
	CMC.T01	1	0.063 ± 0.002	0.013 ± 0.0010
	BRV.Hapo.T01	1	0.039 ± 0.002	0.007 ± 0.0008
	CMC.Hapo.T01	1	0.037 ± 0.002	0.007 ± 0.0007
	VT.BRV.Hapo	1	0.043 ± 0.002	0.008 ± 0.0008
	VT.CMC.Hapo	1	0.040 ± 0.002	0.007 ± 0.0007
	WSS.Hapo	1	0.021 ± 0.001	0.003 ± 0.0005
	PedCMC	1	0.064 ± 0.002	0.014 ± 0.001
	FPCA	1	0.052 ± 0.002	0.012 ± 0.001
	$FARVAT_b$	1	0.049 ± 0.02	0.009 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.052 ± 0.02	0.010 ± 0.001
	$FARVAT_c$	1	0.049 ± 0.02	0.009 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.050 ± 0.02	0.010 ± 0.001
	$FARVAT_o$	1	0.054 ± 0.02	0.011 ± 0.001
		$[p_k(1-p_k)]^{-1/2}$	0.052 ± 0.02	0.011 ± 0.001

If the large number of rare variants is collapsed, its convergence to the chi-square distribution requires very large sample sizes, and genotype scores for all individuals can be 1 in extreme scenarios. Therefore, we could conclude that PedCMC may not be a good choice when the number of rare variants in a gene is very large.

The statistical efficiency of $FARVAT_b$, $FARVAT_c$, and $FARVAT_o$ was evaluated with the simulated data, and results were compared with results from PedCMC, FPCA, and RV-TDT methods [He, et al. 2014; Zhu and Xiong 2012]. We assumed that the relative proportion of variances explained by rare variants h_a^2 was 0.05. In each replicate, we assumed that all rare variants had either deleterious or protective effects on disease, and the proportions of rare variants with deleterious effects were assumed to be 1, 0.8, 0.6, and 0.5. The numbers of extended families were assumed to be 50 and 100. MAFs for all rare variants were assumed to be less than 0.01. Thirty rare variants in Figure 3.2 and 100 rare variants in Figure 3.3 were selected, and in Figure 3.4, all rare variants within 30 kbp from the generated 1 Mbp chromosomes were selected. For the proposed methods, each rare variant was weighed by $[p_k(1 - p_k)]^{-1/2}$ for $\mathbf{W}$. The results in Figures 3.2 – 3.4 showed that $FARVAT_b$ was the most efficient if all rare variants had deleterious effects, but the gap between $FARVAT_b$ and the second efficient method $FARVAT_o$ was small. However, the power loss of $FARVAT_b$ was substantial when rare variants with deleterious and protective variants were present in a gene.

If the proportion of rare variants with deleterious effects was 0.5, $FARVAT_c$ was the most efficient, followed by $FARVAT_o$. PedCMC and FPCA were usually more efficient than RV-TDT methods, but these approaches were not efficient compared to $FARVAT_o$ in the considered scenarios. Therefore, even though the most powerful statistic depended on the disease model, we concluded that $FARVAT_o$ was generally efficient choice under the various disease models.

Figure 3.2 Empirical power estimates when the number of rare variants in a gene is 30. h_a^2 was assumed to be 0.05 and the empirical power estimates were calculated with 5,000 replicates at the 0.001 significance levels. MAFs for all variants were assumed to be less than 0.01, and 30 rare variants were randomly selected. Each rare variant had either deleterious or protective effect on disease, and proportions of rare variants with deleterious effect were 1, 0.8, 0.6 and 0.5. The numbers of families were assumed to be 50 and 100.

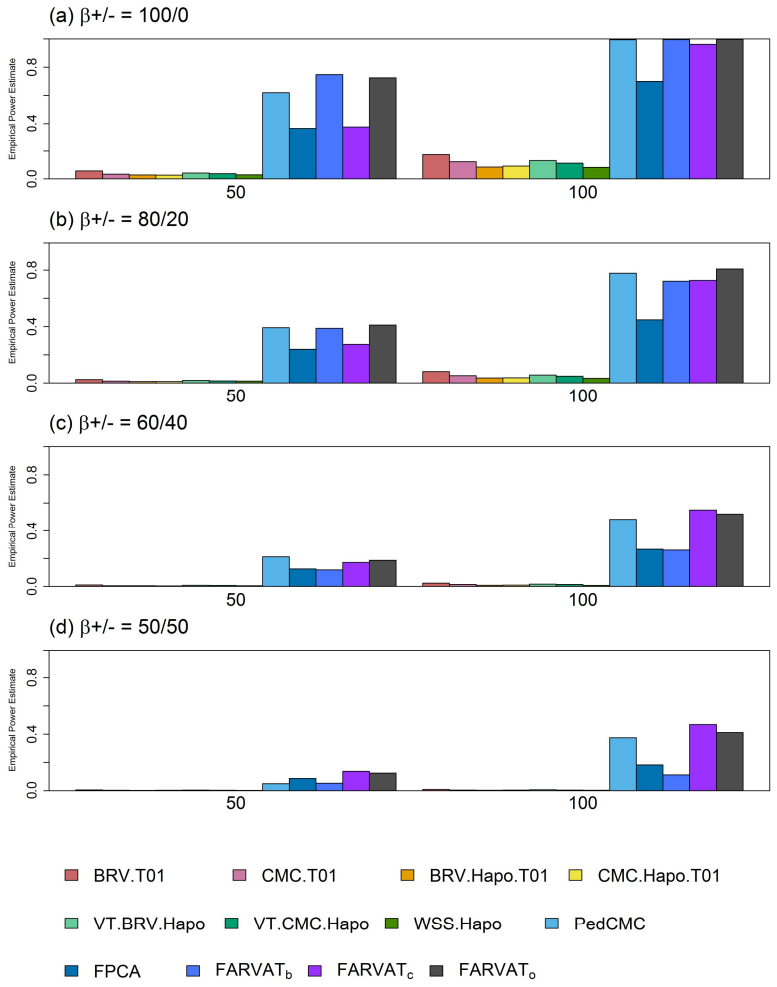


Figure 3.3 Empirical power estimates when the number of rare variants in a gene is 100. h_a^2 was assumed to be 0.05 and the empirical power estimates were calculated with 5,000 replicates at the 0.001 significance levels. MAFs for all variants were assumed to be less than 0.01, and 100 rare variants were randomly selected. Each rare variant had either deleterious or protective effect on disease, and proportions of rare variants with deleterious effect were 1, 0.8, 0.6 and 0.5. The numbers of families were assumed to be 50 and 100.

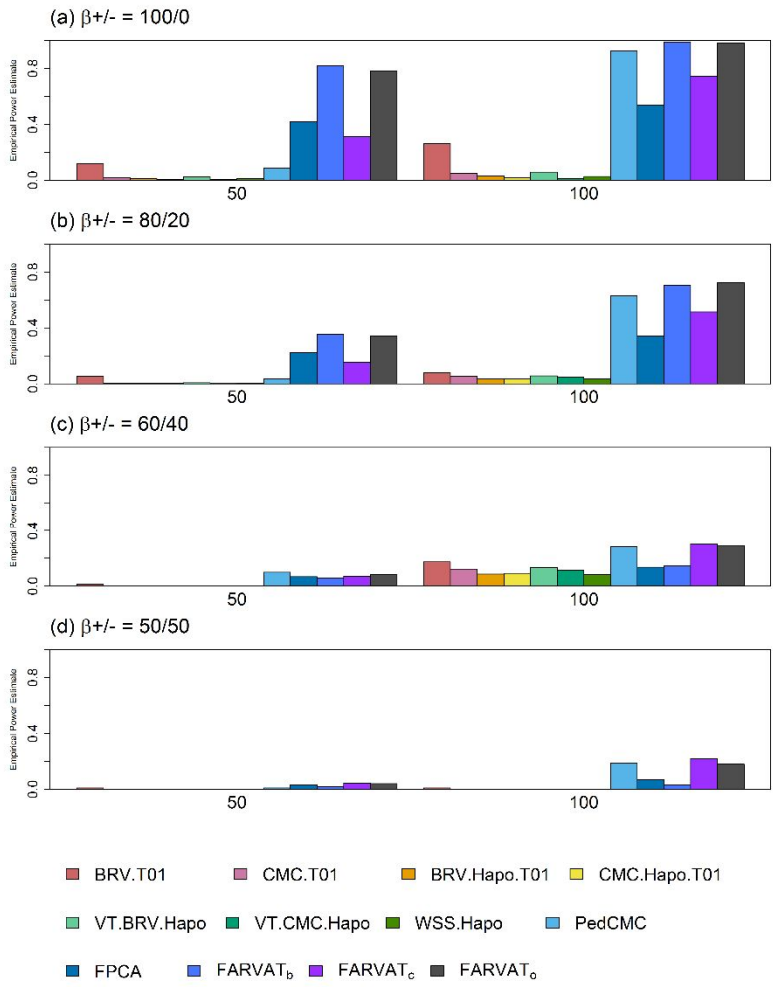
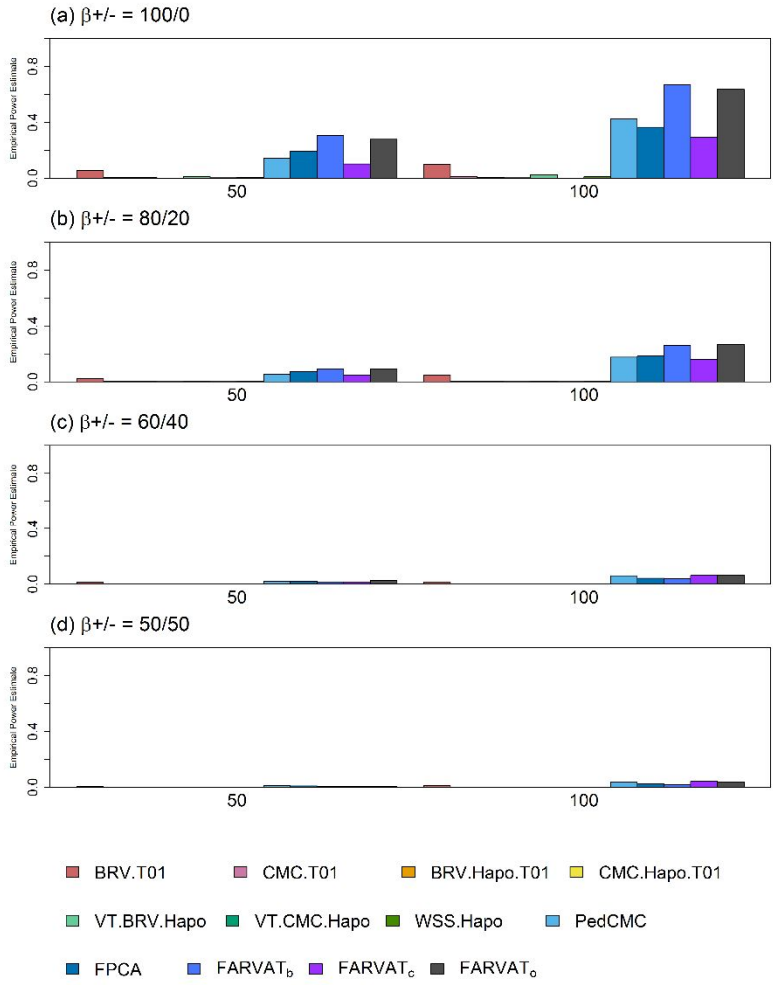


Figure 3.4 Empirical power estimates when all rare variants in a gene are considered. h_a^2 was assumed to be 0.05 and the empirical power estimates were calculated with 5,000 replicates at the 0.001 significance levels. All rare variants of which MAFs are less than 0.01 are used to calculate each statistic. Each rare variant had either deleterious or protective effect on disease, and proportions of rare variants with deleterious effect were 1, 0.8, 0.6 and 0.5. The numbers of families were assumed to be 50 and 100.



3.3.3 Evaluation with simulated data under the presence of population substructure

We assumed that there were two subpopulations, and founders in each family were randomly selected from two different population substructures. Two subpopulations were simulated with the Balding-Nichols model [Balding and Nichols 1995], and F_{ST} values were assumed to be 0.005, 0.01, and 0.05. To provide robustness against the population substructure, $\mathbf{\Phi}$ was estimated by 20,000 common variants for which MAFs were larger than 0.05 [Thornton and McPeek 2010], and this was incorporated to the proposed methods. It should be noted that the proposed method was an extension of the M_{QLS} statistic [Thornton and McPeek 2007] to rare variant association analysis, and that M_{QLS} becomes robust under the presence of population substructure if $\mathbf{\Phi}$ was estimated with large-scale genomic data [Thornton and McPeek 2010].

In Table 3.4, we calculated empirical type-1 error estimates from 50,000 replicates at the 0.05 and 0.01 significance levels. Our results showed that the empirical type-1 error estimates for $FARVAT_b$ and $FARVAT_o$ preserved the nominal significance levels for the considered F_{ST} values. However, FPCA, PedCMC, and RV-TDT were usually conservative, and the level of conservativeness was proportional to the amount of F_{ST}. $FARVAT_c$ was also conservative, but was less sensitive than FPCA, PedCMC, and RV-TDT. Furthermore, we evaluated the statistical efficiency under the presence of population substructure with the simulated data. We assumed that h_a^2 was 0.05 and the empirical power estimates were calculated with 5,000 replicates at the

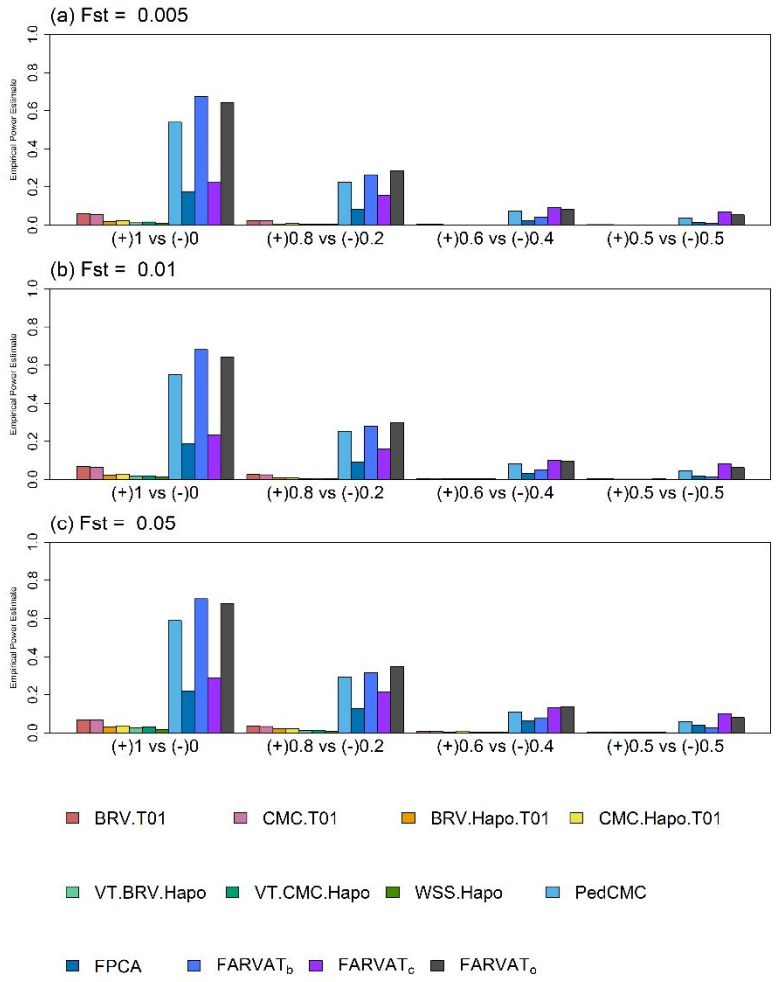
0.001 significance level. As shown in Figure 3.5, the most efficient approach differed depending on the disease model. For instance, $FARVAT_b$ was the most efficient when all rare variants had deleterious effects, and $FARVAT_c$ was the most efficient when half of the rare variants had deleterious effects. $FARVAT_o$ was usually the second most efficient; however, the power gap between $FARVAT_o$ and the most efficient method was always small. As a result, we concluded that $FARVAT_o$ was generally a robust and efficient choice for various disease models under the presence of population substructure.

Table 3.4 Empirical type-1 error estimates under the presence of population substructure. The empirical type-1 error estimates were calculated with 50,000 replicates at the 0.05 and 0.01 significance levels under the presence of population substructure. F_{ST} was assumed to be 0.005, 0.01 and 0.05. MAFs for all variants are assumed to be less than 0.01, and 100 rare variants are randomly selected. The number of families in each replicate is assumed to be 100.

F_{ST}	statistics	weight	α=.05(95%CI)	α=.01(95%CI)
0.005	BRV.T01	1	0.036 ± 0.002	0.006 ± 0.0007
	CMC.T01	1	0.035 ± 0.002	0.005 ± 0.0006
	BRV.Hapo.T01	1	0.018 ± 0.001	0.003 ± 0.0004
	CMC.Hapo.T01	1	0.021 ± 0.001	0.003 ± 0.0005
	VT.BRV.Hapo	1	0.023 ± 0.001	0.004 ± 0.0005
	VT.CMC.Hapo	1	0.024 ± 0.001	0.004 ± 0.0006
	WSS.Hapo	1	0.008 ± 0.001	0.001 ± 0.0003
	PedCMC	1	0.024 ± 0.001	0.005 ± 0.001
	FPCA	1	0.021 ± 0.001	0.004 ± 0.001
	$FARVAT_b$	$[p_k(1-p_k)]^{-1/2}$	0.052 ± 0.002	0.011 ± 0.001
	$FARVAT_c$	$[p_k(1-p_k)]^{-1/2}$	0.039 ± 0.002	0.006 ± 0.001
	$FARVAT_o$	$[p_k(1-p_k)]^{-1/2}$	0.044 ± 0.002	0.008 ± 0.001
0.01	BRV.T01	1	0.036 ± 0.002	0.0063 ± 0.0007
	CMC.T01	1	0.035 ± 0.002	0.0059 ± 0.0007
	BRV.Hapo.T01	1	0.021 ± 0.001	0.0030 ± 0.0005
	CMC.Hapo.T01	1	0.024 ± 0.001	0.0037 ± 0.0005
	VT.BRV.Hapo	1	0.025 ± 0.001	0.0040 ± 0.0006
	VT.CMC.Hapo	1	0.028 ± 0.001	0.0046 ± 0.0006
	WSS.Hapo	1	0.010 ± 0.001	0.0015 ± 0.0003
	PedCMC	1	0.021 ± 0.001	0.004 ± 0.001
	FPCA	1	0.021 ± 0.001	0.004 ± 0.001
	$FARVAT_b$	$[p_k(1-p_k)]^{-1/2}$	0.051 ± 0.002	0.010 ± 0.001
	$FARVAT_c$	$[p_k(1-p_k)]^{-1/2}$	0.038 ± 0.002	0.005 ± 0.001
	$FARVAT_o$	$[p_k(1-p_k)]^{-1/2}$	0.044 ± 0.002	0.007 ± 0.001

0.05	BRV.T01	1	0.042 ± 0.002	0.0075 ± 0.0008
	CMC.T01	1	0.041 ± 0.002	0.0071 ± 0.0007
	BRV.Hapo.T01	1	0.031 ± 0.002	0.0052 ± 0.0006
	CMC.Hapo.T01	1	0.033 ± 0.002	0.0058 ± 0.0007
	VT.BRV.Hapo	1	0.033 ± 0.002	0.0060 ± 0.0007
	VT.CMC.Hapo	1	0.034 ± 0.002	0.0061 ± 0.0007
	WSS.Hapo	1	0.018 ± 0.001	0.0028 ± 0.0005
	PedCMC	1	0.017 ± 0.001	0.003 ± 0.000
	FPCA	1	0.026 ± 0.001	0.005 ± 0.001
	$FARVAT_b$	$[p_k(1-p_k)]^{-1/2}$	0.055 ± 0.002	0.012 ± 0.001
	$FARVAT_c$	$[p_k(1-p_k)]^{-1/2}$	0.042 ± 0.002	0.006 ± 0.001
	$FARVAT_o$	$[p_k(1-p_k)]^{-1/2}$	0.046 ± 0.002	0.009 ± 0.001

Figure 3.5 Empirical power estimates under the presence of population substructure. h_a^2 was assumed to be 0.05 and the empirical power estimates were calculated with 5,000 replicates at the 0.001 significance levels under the presence of population substructure. F_{ST} was assumed to be 0.005, 0.01 and 0.05. MAFs for all variants are assumed to be less than 0.01, and 30 rare variants are randomly selected. Each rare variant had either deleterious or protective effect on disease, and proportions of rare variants with deleterious effect were 1, 0.8, 0.6 and 0.5. The numbers of families were assumed to be 100.



3.3.4 Analysis of GAW17 simulated data

The statistical efficiency of the proposed methods was evaluated with the binary trait in GAW17 simulated data [Almasy, et al. 2011]. There were 200 replicates in GAW17 simulated data, and each replicate consisted of 209 affected and 488 unaffected individuals distributed in eight extended pedigrees. In 1,714 genes, there were 13,784 variants, and MAFs for 10,710 variants were less than 0.05. In each gene, rare variants for which MAFs were less than 0.05 were considered for analysis with the proposed methods, and genes in which the number of rare variants was less than or equal to 2 were excluded from the analysis. To provide the robustness of the proposed methods under the presence of population substructure, the empirical genetic relationship matrix between individuals was estimated with the common variants. The disease status was decided by the underlying liability, and the top 30% of the underlying liability distribution was declared as being affected. In particular, some covariates were related to the underlying liability, and the disease prevalence [Thornton and McPeek 2010] and BLUP from the linear mixed model [Won and Lange 2013] were utilized as offsets. For the linear mixed model, we included sex, age, smoking status, and 10 principal component scores calculated from the estimated $\mathbf{\Phi}$ [Thornton and McPeek 2010]. Among 36 genes related to binary traits, 20 genes consisted of more than one rare variant, and their empirical powers were determined by counting the number of replicates for which p-values of causal genes were smaller than 0.05 and 0.01. As shown in Tables 3.5 and 3.6, most causal genes were not

detectable with the proposed methods; however, *KDR*, *VEGFA*, *SIRT1,* and *VLDLR* had relatively high rates of detection. By using RV-TDT methods, we could not find any causal genes. In Figures 3.6 and 3.7, we provided the qq-plots and Manhattan plots of RV-TDT, PedCMC, $FARVAT_b$, $FARVAT_c$, and $FARVAT_o$ with the first replicate of GAW17 simulated data. While PedCMC was not conservative, results from the other methods seemed to be valid. As shown in Figure 3.7, we found that *VEGFA* was the most significant for $FARVAT_o$.

Table 3.5 Rare variant association analysis with GAW17 simulated data for AFFECTED trait. The numbers of replicates among 200 replicates where *p*-values for each method were less than 0.05 and 0.01 are counted. The correlation matrix was used empirical matrix between individuals was estimated with the common variants. The number of rare variants and true casual variants are denoted by *m* and *c*.

			PedCMC		$FARVAT_b$		$FARVAT_c$		$FARVAT_o$	
GENE	***m***	***c***	**<.05**	**<.01**	**<.05**	**<.01**	**<.05**	**<.01**	**<.05**	**<.01**
ARNT	4	2	0	0	8	0	4	0	4	0
BCHE	9	4	0	0	6	0	11	1	4	1
BCL2L11	4	1	0	0	3	0	6	0	3	0
ELAVL4	5	1	5	0	16	6	5	0	13	5
FLT1	8	5	1	0	5	1	3	0	4	0
HIF3A	6	1	0	0	1	0	1	1	1	1
HSP90AA1	5	1	1	0	2	1	10	2	6	1
KDR	3	4	18	7	38	12	16	4	20	9
LPL	11	2	0	0	5	0	0	0	1	0
PDGFD	6	1	1	0	6	1	5	0	6	1
PIK3C2B	23	4	1	0	9	2	11	3	10	2
PLAT	10	2	5	0	3	0	14	2	8	1
RRAS	3	1	0	0	8	0	2	2	4	1
SIRT1	11	4	3	2	49	14	100	56	81	52

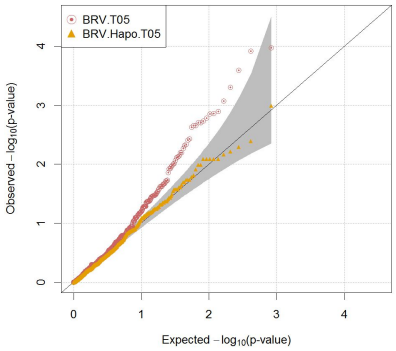
SREBF1	7	5	0	0	6	0	8	1	5	0
VEGFA	3	1	88	40	123	68	135	83	127	89
VLDLR	10	2	10	3	52	19	29	6	51	11
VNN1	3	1	2	0	5	0	5	0	4	0
VNN3	4	3	1	0	3	0	2	0	1	0
VWE	6	2	2	0	0	0	1	0	1	0

Table 3.6 Rare variant results of GAW17 data for AFFECTED trait adjusting for covariates. In 200 replicates, we counted the number of times that *p*-values of proposed method were less than 0.05, 0.01 and 0.001.The correlation matrix was used empirical matrix between individuals was estimated with the common variants. The number of rare variants and casual variants are denoted by *m* and *c*. Covariates included age, sex, smoking status and principal components.

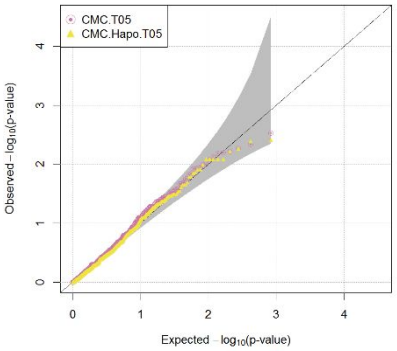
			$FARVAT_b$		$FARVAT_c$		$FARVAT_o$	
GENE	***m***	***c***	**<0.05**	**<0.01**	**<0.05**	**<0.01**	**<0.05**	**<0.01**
ARNT	4	2	17	7	21	4	21	7
BCHE	9	4	18	0	21	6	18	6
BCL2L11	4	1	16	6	19	1	17	2
ELAVL4	5	1	24	7	23	4	28	9
FLT1	8	5	53	23	48	14	47	21
HIF3A	6	1	5	1	7	1	5	1
HSP90AA1	5	1	16	3	0	0	0	0
KDR	3	4	20	5	12	1	19	4
LPL	11	2	12	6	16	4	16	5
PDGFD	6	1	13	2	14	2	13	2
PIK3C2B	23	4	15	2	17	5	18	4
PLAT	10	2	10	4	8	0	8	4
RRAS	3	1	25	3	0	0	0	0
SIRT1	11	4	21	9	19	6	19	10
SREBF1	7	5	33	13	28	7	31	15
VEGFA	3	1	65	29	69	29	78	40
VLDLR	10	2	16	2	5	2	10	2
VNN1	3	1	30	16	26	11	27	16
VNN3	4	3	18	5	17	2	16	3
VWF	6	2	0	0	1	0	1	0

Figure 3.6 QQ-plot of the rare variant association analysis with GAW17 simulated data for AFFECTED trait. The qq-plots are provided for BRV.T05, BRV.Hapo.T05, CMC.T05, CMC.Hapo.T05, VT.BRV.Hapo, VT.CMC.Hapo, WSS.Hapo, PedCMC, $FARVAT_b$, $FARVAT_c$ and $FARVAT_o$. The 95% confidence interval is provided.

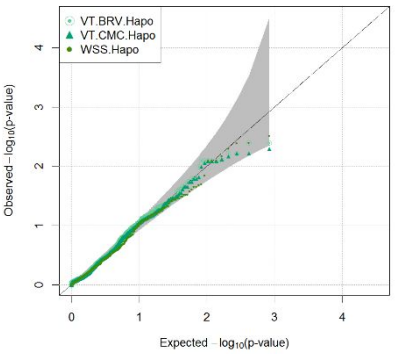
(a) BRV.T05 & BRV.Hapo.T05



(b) CMC.T05 & CMC.Hapo.T05



(c) VT.BRV.Hapo, VT.CMC.Hapo and WSS.Hapo



(d) PedCMC, $FARVAT_b$, $FARVAT_c$ and $FARVAT_o$

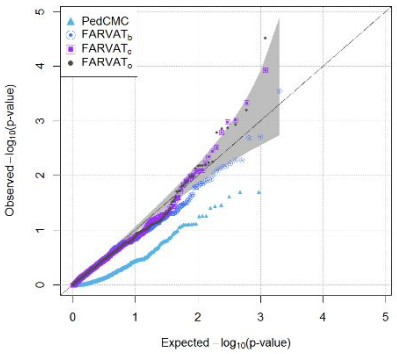
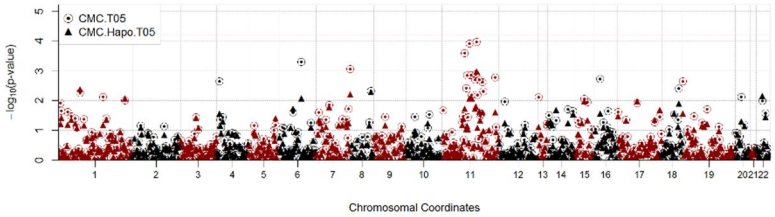
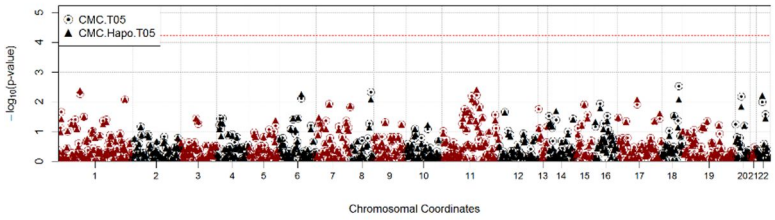


Figure 3.7 Manhattan plot of the rare variant association analysis with GAW17 simulated data for AFFECTED trait. The Manhattan plots are provided for BRV.T05, BRV.Hapo.T05, CMC.T05, CMC.Hapo.T05, VT.BRV.Hapo, VT.CMC.Hapo, WSS.Hapo, PedCMC, $FARVAT_b$, $FARVAT_c$ and $FARVAT_o$. The x-axis indicates the genome in physical position and y-axis does $-\log_{10}(p\text{-value})$ for all genes. The horizontal line means the threshold for 0.05 genome-wide significance level by Bonferroni correction is 2.50E-05.

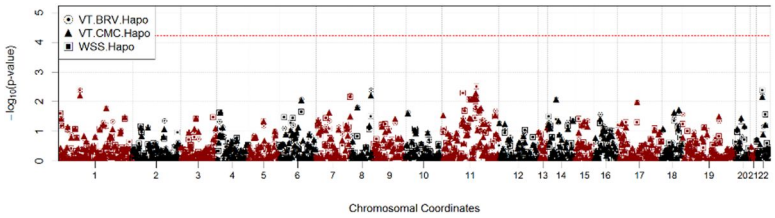
(a) BRV.T05 & BRV.Hapo.T05



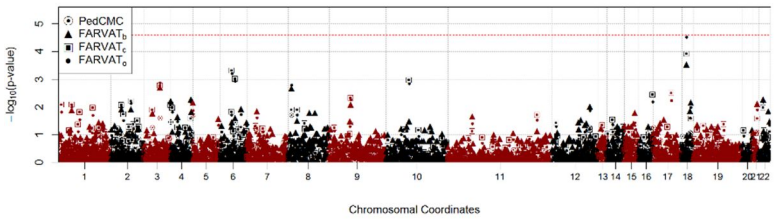
(b) CMC.T05 & CMC.Hapo.T05



(c) VT.BRV.Hapo, VT.CMC.Hapo, and WSS.Hapo



(d) PedCMC, $FARVAT_b$, $FARVAT_c$, and $FARVAT_o$



3.4 Application to schizophrenia data

The proposed methods were applied to the genetic association analysis of rare variants in schizophrenia. Thirty-six trios were collected from Germany for which offspring were affected while parents were unaffected. The whole genomes for all individuals were sequenced. There were 10,829,265 bi-allelic variants, and MAFs of 31,860 among them were less than 0.05. Markers with high missing call rates (> 5%) or significant deviation from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-5}$) were excluded, and trios were filtered out if 10% of variants had Mendelian transmission errors. As a result, 9,216,373 common variants and 31,046 rare variants for 105 trios were analyzed with the proposed methods.

Each rare variant was annotated by the SnpEff program [Cingolani, et al. 2012] with the UCSC HG19 database. SnpEff 3.2a categorized each variant to four groups: HIGH, MODERATE, LOW, and MODIFIER. In our analysis, rare variants assigned to LOW and MODIFIER categories may have little or no effect on protein function, and they were not considered in our analysis. For each gene, the rare variants with HIGH and MODERATE effects were separately analyzed with the proposed methods. In addition, if MAC of all rare variants in each gene were less than or equal to 5, the asymptotic convergence of the proposed method to chi-square distribution may not be provided, and *p*-values were calculated for genes for which the MAC was larger than or equal to 5. In total, *p*-values were calculated for 13,053 genes. For the proposed methods, the prevalence of schizophrenia was assumed to be

0.0063, and each rare variant was weighted by $[p_k(1 - p_k)]^{-1/2}$ for **W**. To provide robustness under the presence of population substructure, the genetic relationship matrix was estimated with common variants, and these data were incorporated into the proposed methods. We provided the qq-plots of RV-TDT methods, PedCMC, $FARVAT_b$, $FARVAT_c$, and $FARVAT_o$. As presented in Figure 3.8, while RV-TDT methods, PedCMC, and $FARVAT_c$ methods were conservative and $FARVAT_b$ showed some violations, $FARVAT_o$ uniquely seems valid. Figure 3.9 shows the Manhattan plots for the all methods, and the genome-wide significant results from RV-TDT, $FARVAT_b$, and $FARVAT_o$ are summarized in Table 3.7. We found a single genome-wide significant gene with WSS.Hapo, $FARVAT_b$, and $FARVAT_o$, and this genome-wide significant gene will be further investigated with replication studies.

Figure 3.8 QQ-plot of the rare variant association analysis for schizophrenia. The qq-plots are provided for BRV.T05, BRV.Hapo.T05, CMC.T05, CMC.Hapo.T05, VT.BRV.Hapo, VT.CMC.Hapo, WSS.Hapo, PedCMC, $FARVAT_b$, $FARVAT_c$ and $FARVAT_o$. The 95% confidence interval is provided.

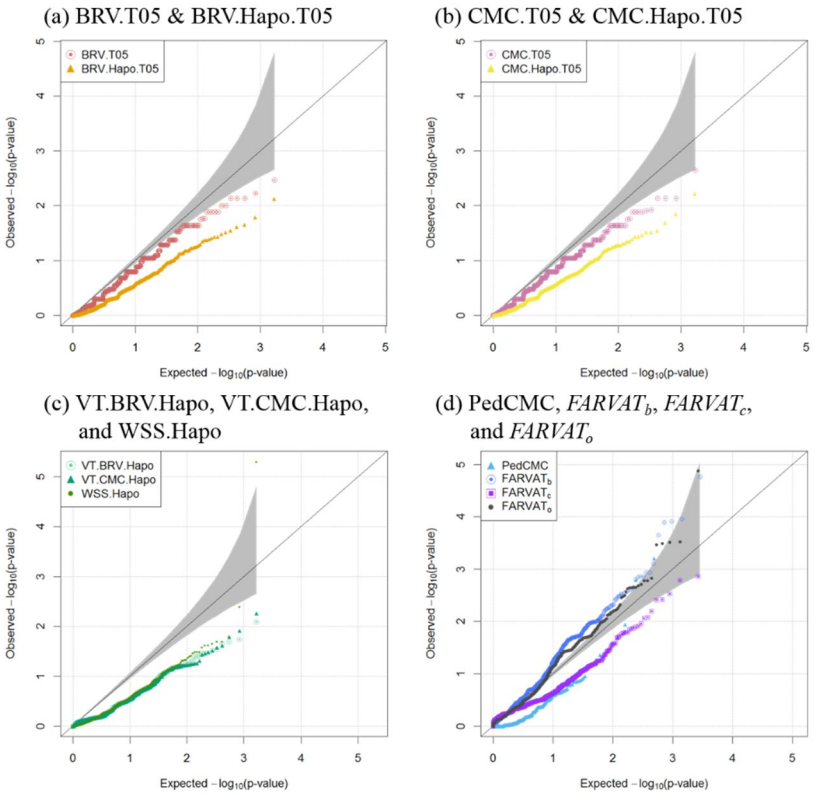


Figure 3.9 Manhattan plot of the rare variant association analysis for schizophrenia. The Manhattan plots are provided for BRV.T05, BRV.Hapo.T05, CMC.T05, CMC.Hapo.T05, VT.BRV.Hapo, VT.CMC.Hapo, WSS.Hapo, PedCMC, $FARVAT_b$, $FARVAT_c$ and $FARVAT_o$. The x-axis indicates the genome in physical position and y-axis does -$\log_{10}$(p-value) for all genes. The horizontal line means the threshold for 0.05 genome-wide significance level by Bonferroni correction is 1.74E-05.

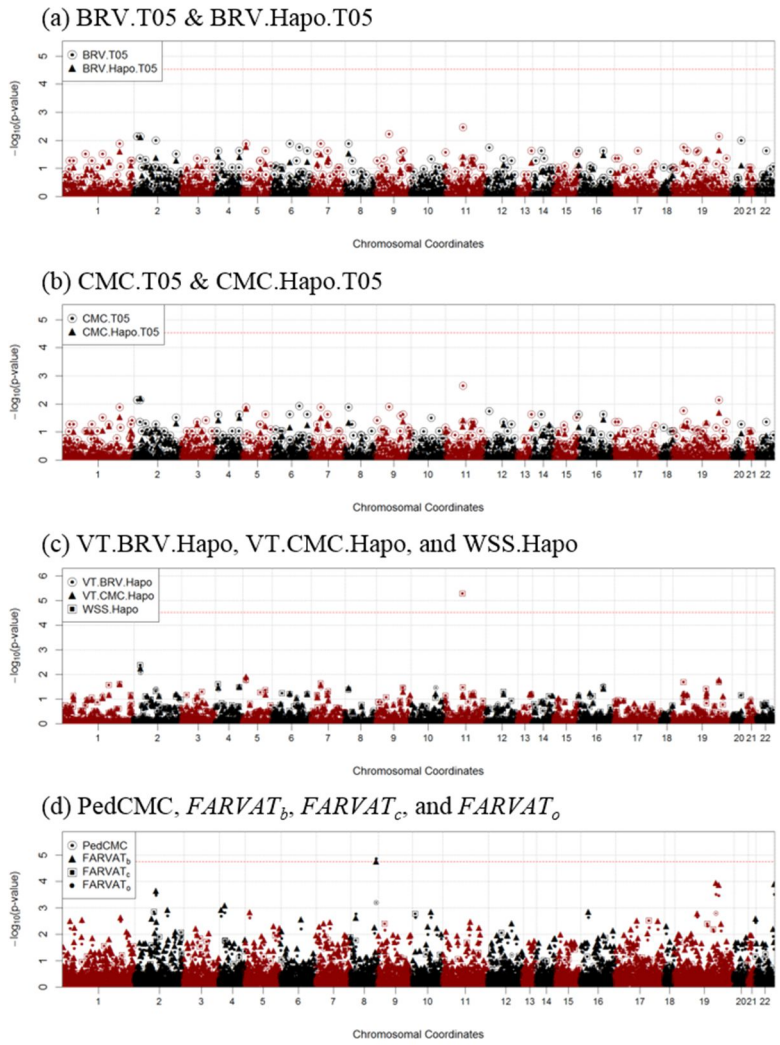


Table 3.7 Significant results from the rare variant association analysis with schizophrenia data.

Statistics	Weight	CHR	GENE	*m*	MAC		*p*-value	*q*-value
					Aff	Unaff		
WSS.Hapo	1	11	Gene1	5	0	11	5.00E-06	0.01
$FARVAT_b$	$[p_k(1-p_k)]^{-1/2}$	8	Gene2	25	4	27	1.67E-05	0.05
$FARVAT_o$	$[p_k(1-p_k)]^{-1/2}$	8	Gene2	25	4	27	1.30E-05	0.04

Notes. The significant results for each method are provided. The numbers of variants for each significant region are provided, and MAC for affected and unaffected individuals is provided. The 0.05 genome-wide significant level adjusted by Bonferroni correction is 1.74E-05, and *q*-values [Benjamini and Hochberg 1995] are provided.

3.5 Discussion

In this chapter, we proposed burden-type, C-alpha-type and SKAT-O-type statistics for the association analysis of rare variants for binary traits with extended families. The proposed methods were compared with results of PedCMC, FPCA, and RV-TDT methods [He, et al. 2014; Zhu and Xiong 2012] and with extensive simulations, we showed that the proposed method was more efficient than existing approaches. In particular, we found that the most efficient statistic among the proposed statistics differed according to the disease model. However, they were usually followed by the SKAT-O-type statistic in such scenarios, and the power differences between the most efficient statistic and the SKAT-O-type statistic were small. Therefore, $FARVAT_o$ seemed to be a robust choice for the analysis of rare variants in extended families.

Furthermore, the proposed method was very rapid computationally, and the *FARVAT* software for the proposed methods was implemented with C++ to enhance computational efficiency. The time complexity for the proposed method was $O(m^3 + N^2m + N^3)$, and we found that analysis of the whole genome sequence data for 1,000 individuals in the extended family design could be conducted within a few hours. *FARVAT* can handle various input file formats, such as the ped, bed, and vcf files, and multithreaded genome-wide association analyses can be conducted. The software calculates various statistics for the analysis of extended families, and it is freely downloadable from http://healthstat.snu.ac.kr/software/farvat/.

However, despite the analytical flexibility of the proposed method, it has some limitations. First, the proposed method could be less efficient if some covariates associated with disease status or phenotypes of interest were continuous. Our recent investigation found that the power improvement of the analysis with phenotypes adjusted by BLUP could be substantial if each family was randomly selected [Won and Lange 2013]. Under certain scenarios, however, power loss may be expected, and the further investigation is necessary. Second, we showed that incorporation of the estimated correlation matrix to the proposed statistics provided sufficient robustness for the proposed method against the presence of population substructure. However, if large-scale common variants were not available or the level of population substructure depended on the genomic location, the proposed adjustment with the estimated correlation matrix did not perform appropriately [Price, et al. 2006; Won, et al. 2009], and different strategies would be necessary according to the level of population substructure. If large-scale common variants are not available, the FBAT or TDT statistics, based on so-called within-family components, is uniquely robust to population substructure, and the burden-type test for the FBAT statistic or RV-TDT can be utilized [De, et al. 2013; He, et al. 2014]. If the genomic ancestry for each individual differs for some genomic locations, the so-called hybrid-analysis strategy [Won, et al. 2009] can be suitable alternative. The proposed method can simply be extended to the statistic based on the between-family component [Won and Lange 2013],

and its rank-based *p*-value can be combined with the FBAT burden type test or SKAT-O-type test.

Advances in genotyping technology will lead to substantial cost reductions for genome sequencing and it is expected that whole genome sequencing may be feasible for less than a few hundred US dollars in the near future. Importantly, most of human genome consists of rare variants, and thus, we expect that the genetic background for 'missing heritability' can be determined by rare variant association analysis [Manolio, et al. 2009]. However, rare variant association analysis is disrupted by genetic heterogeneity, and in this context, the importance of rare variant analysis with extended families has often been raised [Ionita-Laza, et al. 2011]. The proposed method enables the analysis of rare variants within extended families, and its application to extended families may provide a breakthrough for the success of genetic association analysis.

This chapter was accepted in *Genetic Epidemiology*
as a partial fulfillment of Sungkyoung Choi's Ph.D program.

Chapter 4

Family-based Rare Variant Association Test for X-linked genes

4.1 Introduction

In this chapter we propose a family-based rare variant association test for X-linked genes (***FARVATX***) that is applicable to various biological models. Due to the nature of our statistic, the proposed method can also be applied to family-based designs with dichotomous phenotype, and we show with extensive simulation studies that the proposed methods perform better than the existing approaches. We applied the coding strategy that was suggested by Wang et al. [Wang, et al. 2014] in population-based design. The proposed methods were applied to an association analysis of families with chronic

obstructive pulmonary disease (COPD). Some promising genes were identified with the proposed methods, thereby illustrating the practical value of these methods.

4.2 Methods

4.2.1 Notation

We assume that there are n families and n_i individuals in family i, and the total sample size is denoted by $N = \sum_{i=1}^{n} n_i$. We assume that genotypes for M rare variants on the X chromosome are available. We let x_{ij}^m be the coded genotype of an individual j in a family i for a variant m, with allowed values of 0, 1, or 2 for a female, and 0 or 1 for a male individual, depending on the number of minor alleles. We denote the disease prevalence by q and assume that y_{ij} is coded as 1 for affected individuals, q for individuals with missing phenotype, and 0 for unaffected individuals. In retrospective analyses, genetic association is detected by comparing genetic distributions of affected and unaffected individuals, and it has been shown that the statistical efficiency can be improved by modifying the phenotype [Lange and Laird 2002; Thornton and McPeek 2007]. We let μ_{ij} be the offset that is define by disease prevalence or the best linear unbiased predictor (BLUP) from the linear mixed model [Won and Lange 2013], and set $t_{ij} = y_{ij} - \mu_{ij}$. Then, if we represent the column vectors that comprise x_{ij}^m and t_{ij} for all individuals in a family i by $\mathbf{X}_i^m$ and $\mathbf{T}_i$ respectively, the genotype matrix and phenotype vector can be defined by

$$\mathbf{X}^m = \begin{pmatrix} \mathbf{X}_1^m \\ \vdots \\ \mathbf{X}_n^m \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} \mathbf{X}^1 & \cdots & \mathbf{X}^M \end{pmatrix}, \text{ and } \mathbf{T} = \begin{pmatrix} \mathbf{T}_1 \\ \vdots \\ \mathbf{T}_n \end{pmatrix}.$$

4.2.2 Variance covariance matrix

We assume that $\sigma_{mm'}$ is a covariance between x_{ij}^m and $x_{ij}^{m'}$ when an individual j in a family i is a male, and the genetic variance-covariance matrix between M markers in males is

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1M} \\ \vdots & \ddots & \vdots \\ \sigma_{M1} & \cdots & \sigma_{MM} \end{pmatrix}.$$

We assume that h_{ij} is an inbreeding coefficient for an individual j in a family i, and thus if an individual j is a male, h_{ij} becomes 0. $\pi_{ij,i'j'}$ is a kinship coefficient between an individual j in a family i and an individual j' in a family i'. It should be noted that $\pi_{ij,i'j'}$ is a function of gender, and can be deductively calculated [Thornton, et al. 2012]. If i and i' are different, $\pi_{ij,i'j'}$ becomes 0. We consider individuals j and j' in a family i, and if an individual j is a descendant of j', $\pi_{ij,ij'}$ can be derived based on Table 4.1.

Table 4.1 X chromosomal and autosomal kinship coefficients for two individuals in a nuclear family.

Relationship of individuals *ij* and *ij′*	$\pi_{ij,ij'}$ (X chromosome)	$\pi_{ij,ij'}$ (autosome)
Brother & Brother	1/4	1/4
Sister & Sister	3/4	1/4
Brother & Sister	1/4	1/4
Mother & Son	1/2	1/4
Mother & Daughter	1/2	1/4
Father & Son	0	1/4
Father & Daughter	1/2	1/4

We consider the case where individual j in a family i is not a descendant of j' in a large family. If we let $m(j')$ and $f(j')$ indicate the mother and father of j', respectively, $\pi_{ij,ij'}$ can be recursively calculated as follows:

1. $\pi_{ij,ij'} = \pi_{ij,im(j')} / 2$, if j' is a male.

2. $\pi_{ij,ij'} = \pi_{ij,im(j')} / 2 + \pi_{ij,if(j')}$, if j' is a female.

If we define $\mathbf{\Phi}$ by

$$\mathbf{\Phi}_{i,i} = \begin{pmatrix} 1+h_{i1} & 2\pi_{i1,i2} & \cdots \\ 2\pi_{i2,i1} & 1+h_{i2} & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}, \ \mathbf{\Phi}_{i,i'} = \begin{pmatrix} 0 & 0 & \cdots \\ 0 & 0 & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix} \ (i \neq i'),$$

$$\text{and } \mathbf{\Phi} = \begin{pmatrix} \mathbf{\Phi}_{1,1} & \mathbf{\Phi}_{1,2} & \cdots \\ \mathbf{\Phi}_{2,1} & \mathbf{\Phi}_{2,2} & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix},$$

then we have var($\mathbf{X}^m$) = $\sigma^2_{mm}\mathbf{\Phi}$.

If we let $\mathbf{l}_N$ be the $N \times 1$ column vector of which the elements are 1 for male and 2 for female, respectively, the best linear unbiased estimator for $E(\mathbf{X})$ under the null hypothesis can be derived, with some modification of the methods of McPeek et al [McPeek, et al. 2004], by

$$\hat{E}(\mathbf{X}) = \mathbf{l}_N \left(\mathbf{l}_N^t \mathbf{\Phi}^{-1} \mathbf{l}_N \right)^{-1} \mathbf{l}_N^t \mathbf{\Phi}^{-1} \mathbf{X},$$

and $\mathbf{\Sigma}$ can be estimated by

$$\hat{\mathbf{\Sigma}} = \frac{1}{(N-1)} \left[\mathbf{X}^t \mathbf{\Phi}^{-1} \mathbf{X} - \left(\mathbf{l}_N^t \mathbf{\Phi}^{-1} \mathbf{l}_N \right)^{-1} \left(\mathbf{l}_N^t \mathbf{\Phi}^{-1} \mathbf{X} \right)^2 \right].$$

4.2.3 Weighted quasi-likelihood score

We assume that $\mathbf{D}_d$ is a $N \times N$ diagonal matrix, and its diagonal elements are 1 or d if the corresponding individuals are males or females, respectively. X-linked gene expression processes are considered by replacing the genotype matrix $\mathbf{X}$ by $\mathbf{D}_d\mathbf{X}$. $\mathbf{D}_d\mathbf{X}$ will be called the weighted quasi-likelihood score in the remainder of this chapter. The efficient choice of d is related to the gene expression process and can be obtained by considering the relative proportion of each genotype's expression [Clayton 2008]. In particular, homozygous disease genotypes are not usually observed for rare variants; thus, an approximately efficient coding strategy can be chosen by comparing gene expression levels for heterozygous disease genotypes in females and hemizygous disease genotypes in males. Therefore under our coding strategy, XCI and escaped XCI (E-XCI) are efficiently tested with $d = 0.5$ and $d = 1$, respectively. We also have considered another simulation scenario for skewed XCI (S-XCI) owing to nonrandom XCI. S-XCI have been defined using an arbitrary threshold as inactivation of deleterious or normal allele in more than 75% cells [Abkowitz, et al. 1998]. We assumed that the value of d was set as 0.75 or 0.25 to represent S-XCI toward to the deleterious allele or the normal allele, respectively.

4.2.4 Rare X-linked variant association tests

The quasi-likelihood-based score [Won and Lange 2013] for $\mathbf{D}_d\mathbf{X}$ can be defined by

$$\mathbf{T}^t\left(\mathbf{D}_d\mathbf{X}-E\left(\mathbf{D}_d\mathbf{X}\right)\right)=\mathbf{T}^t\mathbf{D}_d\left(\mathbf{X}-E\left(\mathbf{X}\right)\right).$$

Because $E(\mathbf{X})$ can be estimated by $\mathbf{l}_N(\mathbf{l}_N^t\mathbf{\Phi}^{-1}\mathbf{l}_N)^{-1}\mathbf{l}_N^t\mathbf{\Phi}^{-1}\mathbf{X}$, the quasi-likelihood score becomes

$$\mathbf{T}^t\mathbf{D}_d\left(\mathbf{I}_N-\mathbf{l}_N\left(\mathbf{l}_N^t\mathbf{\Phi}^{-1}\mathbf{l}_N\right)^{-1}\mathbf{l}_N^t\mathbf{\Phi}^{-1}\right)\mathbf{X}=\mathbf{T}^t\mathbf{D}_d\mathbf{\Phi}\mathbf{P}\mathbf{X}$$

$$\text{where}\quad \mathbf{P}=\mathbf{\Phi}^{-1}-\mathbf{\Phi}^{-1}\mathbf{l}_N\left(\mathbf{l}_N^t\mathbf{\Phi}^{-1}\mathbf{l}_N\right)^{-1}\mathbf{l}_N^t\mathbf{\Phi}^{-1}.$$

If we let $\mathbf{H} = \mathbf{\Phi} - \mathbf{l}_N(\mathbf{l}_N^t\mathbf{\Phi}^{-1}\mathbf{l}_N)^{-1}\mathbf{l}_N^t$, we can simply show that

$$\operatorname{cov}\left(\mathbf{T}^t\mathbf{D}_d\mathbf{\Phi}\mathbf{P}\mathbf{X}^m,\mathbf{T}^t\mathbf{D}_{d'}\mathbf{\Phi}\mathbf{P}\mathbf{X}^{m'}\right)=\left(\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}\right)\sigma_{mm'},$$

and thus we have

$$\operatorname{cov}\left(\left(\mathbf{T}^t\mathbf{D}_d\mathbf{\Phi}\mathbf{P}\mathbf{X}\right)^t,\left(\mathbf{T}^t\mathbf{D}_{d'}\mathbf{\Phi}\mathbf{P}\mathbf{X}\right)\right)=\left(\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}\right)\mathbf{\Sigma}.$$

It has been empirically shown that weighting each variant can be an efficient strategy to improve statistical power for rare variant association analyses [Madsen and Browning 2009]. We let the weight for variant m be w_m, and the diagonal matrix for which the diagonal element m is w_m be $\mathbf{W}$. If we let p_m be the MAF for a variant m, we used $Beta(p_m; 1, 25)$ as w_m. Then scores for burden [Li and Leal 2008] and variance component [Neale, et al. 2011; Wu, et al. 2011] tests can be respectively defined by

$$\frac{1}{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\mathbf{T}^t\mathbf{D}_d\mathbf{\Phi}\mathbf{P}\mathbf{X}\mathbf{W}\left(0\cdot\mathbf{I}_M+(1-0)\cdot\mathbf{1}_M\mathbf{1}_M^t\right)\mathbf{W}\mathbf{X}^t\mathbf{P}\mathbf{\Phi}\mathbf{D}_d\mathbf{T},$$

and $$\frac{1}{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\mathbf{T}^t\mathbf{D}_d\mathbf{\Phi}\mathbf{P}\mathbf{X}\mathbf{W}\left(1\cdot\mathbf{I}_M+(1-1)\cdot\mathbf{1}_M\mathbf{1}_M^t\right)\mathbf{W}\mathbf{X}^t\mathbf{P}\mathbf{\Phi}\mathbf{D}_d\mathbf{T}.$$

These are extensions of FARVAT statistics [Choi, et al. 2014]. We let $\mathbf{R}_c = c\mathbf{I}_M + (1\text{-}c)\mathbf{1}_\mathrm{M}\mathbf{1}_M^t$ and define

$$S_c^{(d)} = \frac{1}{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\mathbf{T}^t\mathbf{D}_d\mathbf{\Phi}\mathbf{P}\mathbf{X}\mathbf{W}\mathbf{R}_c\mathbf{W}\mathbf{X}^t\mathbf{P}\mathbf{\Phi}\mathbf{D}_d\mathbf{T}.$$

We let *p*-values for $S_c^{(d)}$ be $P_c^{(d)}$, and denote $P_0^{(d)}$ and $P_1^{(d)}$ by ***FARVAT-XB*$_{(d)}$** and ***FARVAT-XC*$_{(d)}$**. It should be noted that the formal corresponds to the burden-type statistic and the latter does SKAT-type statistic. The SKAT-O-type statistic [Lee, et al. 2012b] can be defined by

$$\min\left\{P_0^{(d)}, P_{0.1}^{(d)}, \ldots, P_1^{(d)}\right\},$$

and we denote its *p*-values by ***FARVAT-XO*$_{(d)}$**. *P*-values can be calculated by the numerical algorithms for ***FARVAT*** statistics [Choi, et al. 2014].

If the biological gene expression processes of X-linked genes are not clear, the proposed statistics may be sensitive to the choice of *d*, and a robust statistic needs to be provided. We calculate ***FARVAT-XB*$_{(d)}$** or ***FARVAT-XC*$_{(d)}$** for various choices of *d*, and then combine them to a single *p*-value by using extended Fisher's method for correlated *p*-values [Brown 1975]. We denote its *p*-value by ***FARVAT-XD*** where 0, 0.05, 0.1, …, 0.95, and 1 were considered for d_1, …, and d_L.

If we let $\mathbf{A}_d = (\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T})^{-1/2}\mathbf{W}\mathbf{X}^t\mathbf{P}\boldsymbol{\Phi}\mathbf{D}_d\mathbf{T}$, the rare variant tests for burden-type and SKAT-type can be expressed as quadratic forms and their quadratic forms for a series of d_1, …, and d_L are denoted by

$$Q_1 = \mathbf{A}_1{}^t\mathbf{A}_1,\ Q_2 = \mathbf{A}_1{}^t\mathbf{1}_M\mathbf{1}_M^t\mathbf{A}_1,\ Q_3 = \mathbf{A}_2{}^t\mathbf{A}_2 ...,\ Q_{2L} = \mathbf{A}_L{}^t\mathbf{1}_M\mathbf{1}_M^t\mathbf{A}_L .$$

Based on the results in the previous section, we can simply show that

$$\mathrm{cov}\left(\mathbf{T}^t\mathbf{D}_d\boldsymbol{\Phi}\mathbf{P}\mathbf{X}^m w_m, \mathbf{T}^t\mathbf{D}_{d'}\boldsymbol{\Phi}\mathbf{P}\mathbf{X}^{m'} w_{m'}\right) = \left(\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}\right)\sigma_{mm'}w_m w_{m'},$$

$$\text{and}\quad \mathrm{cov}\left(\left(\mathbf{T}^t\mathbf{D}_d\boldsymbol{\Phi}\mathbf{P}\mathbf{X}\mathbf{W}\right)^t, \mathbf{T}^t\mathbf{D}_{d'}\boldsymbol{\Phi}\mathbf{P}\mathbf{X}\mathbf{W}\right) = \left(\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}\right)\mathbf{W}\boldsymbol{\Sigma}\mathbf{W},$$

which indicates

$$\mathrm{cov}\left(\mathbf{A}_d, \mathbf{A}_{d'}^t\right) = \frac{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}{\sqrt{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\sqrt{\mathbf{T}^t\mathbf{D}_{d'}\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}}\mathbf{W}\boldsymbol{\Sigma}\mathbf{W} \quad \text{and}$$

$$\mathrm{cov}\left(\mathbf{1}_M^t\mathbf{A}_d, \mathbf{A}_{d'}^t\right) = \frac{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}{\sqrt{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\sqrt{\mathbf{T}^t\mathbf{D}_{d'}\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}}\mathbf{1}_M^t\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}.$$

Therefore, under asymptotic normality, covariances between the quadratic forms become

$$\sigma_{Q_l Q_{l'}} = \begin{cases} \dfrac{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}{\sqrt{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\sqrt{\mathbf{T}^t\mathbf{D}_{d'}\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}}\mathrm{tr}\left\{\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^2\boldsymbol{\Sigma}\mathbf{W}\right\} & \text{if } l, l' : \text{odd} \\ \dfrac{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}{\sqrt{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\sqrt{\mathbf{T}^t\mathbf{D}_{d'}\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}}\left(\mathbf{1}_M^t\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\mathbf{1}_M\right)^2 & \text{if } l, l' : \text{even} \\ \dfrac{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}{\sqrt{\mathbf{T}^t\mathbf{D}_d\mathbf{H}\mathbf{D}_d\mathbf{T}}\sqrt{\mathbf{T}^t\mathbf{D}_{d'}\mathbf{H}\mathbf{D}_{d'}\mathbf{T}}}\mathbf{1}_M^t\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^2\boldsymbol{\Sigma}\mathbf{W}\mathbf{1}_M & \text{if } l : \text{odd}, l' : \text{even} \end{cases},$$

and the variances of the quadratic forms are

$$\sigma_{Q_l Q_l} = \begin{cases} \mathrm{tr}\left\{\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^2\boldsymbol{\Sigma}\mathbf{W}\right\} & \text{if } l : \text{odd number} \\ \left(\mathbf{1}_M^t \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\mathbf{1}_M\right)^2 & \text{if } l : \text{even number.} \end{cases}$$

Therefore the correlation between the quadratic forms can be calculated, and they will be denoted as $\rho_{Q_l Q_l}$. If we denote *p*-values for the quadratic form Q_l by p_{Q_l}, we consider

$$S = -2\sum_{l=1}^{2L} \log p_{Q_l} .$$

Here p_{Q_l} can be calculated by the numerical algorithm developed by Davis [Davies 1980a]. Under the null hypothesis, the variance of *S* can be obtained by

$$E\left(S^2\right) = 4L, \;\; \mathrm{var}\left(S\right) = 8L + 2\sum_{l}\sum_{l'<l} \mathrm{cov}\left(-2\log p_{Q_l}, -2\log p_{Q_{l'}}\right).$$

As was suggested by Brown [Brown 1975], the covariance can be approximated by

$$\mathrm{cov}\left(-2\log p_{Q_l}, -2\log p_{Q_t}\right) \approx 3.279\rho_{Q_l Q_t} + 0.711\rho_{Q_l Q_t}^2$$
$$\mathrm{cov}\left(-2\log p_{Q_l}, -2\log p_{Q_t}\right) \approx 3.263\rho_{Q_l Q_t} + 0.710\rho_{Q_l Q_t}^2 + 0.027\rho_{Q_l Q_t}^3 .$$

Under the null hypothesis, *S* is approximately equal to the scaled chi-square distribution as follows:

$$S = -2\sum_{l=1}^{2L} \log p_{Q_l} \sim c\chi^2(\mathrm{df} = f) .$$

Here *c* and *f* can be derived as

$$f = 2\frac{E\left(S^2\right)^2}{\mathrm{var}\left(S\right)},\ c = \frac{\mathrm{var}\left(S\right)}{2E\left(S^2\right)}.$$

4.3 Simulation study

4.3.1 The simulation model

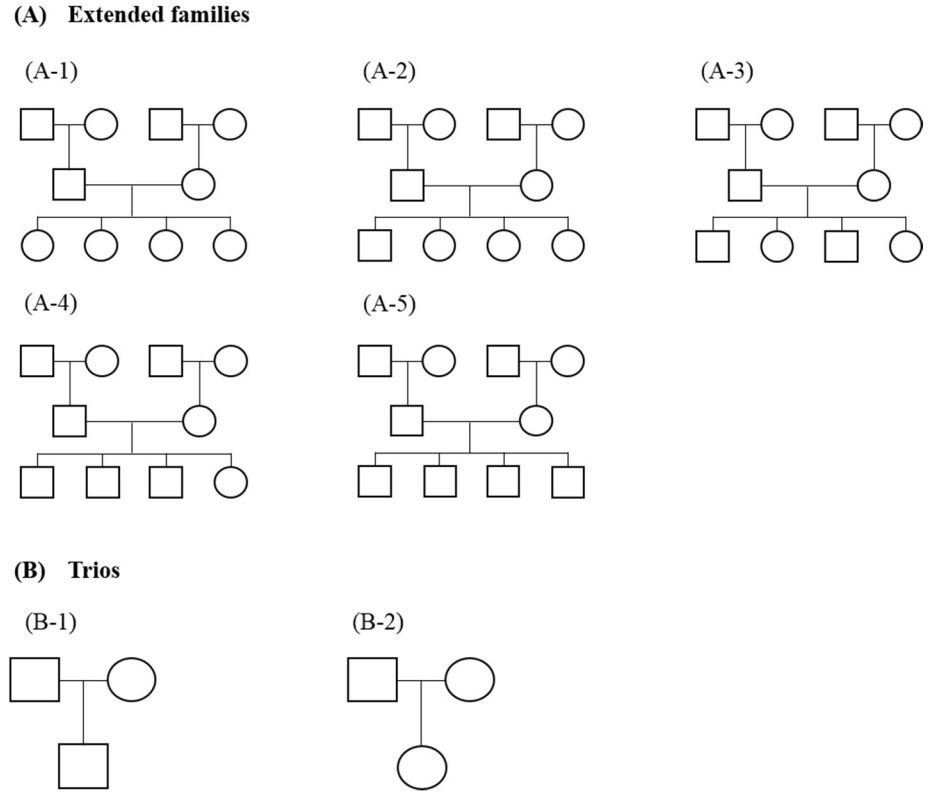
To investigate the performance of the proposed methods, we performed simulation studies for various family structures (see Figure 4.1 for detailed information). We considered trios with a son or a daughter, and large families with 10 individuals that extended over three generations and had different numbers of males and females. MAFs were generated from a uniform distribution $U(0,\ 0.01)$, and genotype frequencies were calculated under Hardy-Weinberg Equilibrium. If we let p_m be the MAF for a variant m, founders' genotypes were generated with a binomial distribution $B(2,\ p_m)$, and offspring's genotypes were obtained by simulated Mendelian transmission, assuming no recombination. Phenotypes for each individual were generated with a liability threshold model, and liabilities were determined by summing the phenotypic mean ($\bar{\mu}$), polygenic effect (σ_g^2), common environmental effect (σ_c^2), main genetic effect and random error (σ_e^2). Random errors were independently generated from $N(0,\ \sigma_e^2 = 1/3)$. The polygenic effect for founders was independently generated from $N(0,\ \sigma_g^2 = 1/3)$, and for non-founders, averages of maternal and paternal polygenic effects were combined with values independently sampled from $N(0,\ 0.5\sigma_g^2)$. Common

environmental effects were assumed to be the same for all individuals in each family and were generated from $N(0, \sigma_c^2 = 1/3)$. For main genetic effects, we assumed that there were M rare variants, and genetic effects for each rare variant were obtained by the product of β_m, the number of disease alleles, and d. If we let h_a^2 be the proportion of phenotypic variance explained by the main genotype, β_m were sampled from $U(1.0, v)$ and v was calculated by

$$v = \sqrt{\frac{\left(\sigma_g^2 + \sigma_c^2 + \sigma_e^2\right) h_a^2}{\left(1 - h_a^2\right) d^2 \sum_{m=1}^{M} \beta_m^2 2 p_m \left(1 - p_m\right)}}.$$

Under the null hypothesis, h_a^2 was set to be 0, and β_m became 0. Liabilities for each individual were generated from the sum of the main genetic effects, polygenic effects, common environmental effects, and random errors, and they were transformed to being affected if they were larger than the threshold; otherwise, they were considered to be unaffected.

Figure 4.1 Family structures considered in our simulation studies



The threshold was set to generate the assumed prevalence q. Disease prevalences are sometimes different between males and females, and this was considered by setting different prevalence rates for males and females in our simulations. Randomly selected families can have very few affected individuals, which leads to the large false negative finding. Therefore, we considered some ascertainment strategies. That is, families with less than two affected grandchildren were excluded from the simulation studies, and sampling was repeated until the desired number of families was obtained.

We also evaluated the proposed methods in the presence of population substructure. We assumed two underlying sub-populations, and each founder was randomly assigned to one of two sub-populations. The polygenic effect, common environmental effect, and random errors were generated with the same model used in the absence of population substructure. However, the phenotypic means of liabilities between two sub-populations were varied by 0.5. The allele frequencies for the two subpopulations were generated with the Balding-Nichols model [Balding and Nichols 1995]. We first generate p_m for global population MAF from $U(0, 0.05)$. Then, if we let F_{ST} denote Wright's F_{ST}, MAFs for two sub-populations were independently sampled from $Beta(p_m(1 - F_{ST})/F_{ST}, (1 - p_m)(1 - F_{ST})/F_{ST})$. F_{ST} was assumed to 0, 0.005, 0.01, and 0.05.

4.3.2 Evaluation with simulated data

We estimated type-1 error rates and powers of the proposed methods, and results from the proposed method were compared with PedGene-Burden and PedGene-Kernel statistics [Schaid, et al. 2013]. In particular, PedGene-Burden and PedGene-Kernel cannot handle S-XCI model and they were not considered for S-XCI model. We considered five different extended family structures (A-1) – (A-5) as shown in Figure 4.1. We assumed that there were 200 extended families and 30 rare variants in each gene. Empirical type-1 errors were calculated at the 0.05 and 0.01 significance levels with 5,000 replicates for dichotomous phenotypes. Tables 4.2 and 4.3 show that type-1 error estimates of our proposed methods consistently preserved the nominal significance levels for any biological expression process, whereas the statistical validity of PedGene-Burden and PedGene-Kernel depends on family structure and type-1 error estimates of PedGene-Burden are violated for (A-1), (A-2), (A-4), and (A-5) of E-XCI. Tables 4.4 – 4.7 show the type-1 error estimates when disease prevalences for males and females are different. Disease prevalences were set to be 0.36 and 0.12 for males and females respectively in Tables 4.4 – 4.5, and 0.12 and 0.36 in Tables 4.6 – 4.7. Results show that the proposed methods always preserve the nominal significance levels. However type-1 error estimates of PedGene-Burden and PedGene-Kernel for E-XCI model setting consistently preserved the nominal significance levels.

Table 4.2 Empirical type-1 error estimates for XCI or E-XCI. Empirical type-1 errors were calculated for five different family structures (A-1) – (A-5). The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01 significance levels.

Type of family	Statistics	Biological Model	α=.05(95%CI)	α=.01(95%CI)
(A-1)	***FARVAT-XB***	XCI	0.0464 ± 0.0058	0.0084 ± 0.0025
		E-XCI	0.0558 ± 0.0064	0.0100 ± 0.0028
	FARVAT-XC	XCI	0.0538 ± 0.0063	0.0098 ± 0.0027
		E-XCI	0.0470 ± 0.0059	0.0104 ± 0.0028
	FARVAT-XO	XCI	0.0500 ± 0.0060	0.0128 ± 0.0031
		E-XCI	0.0578 ± 0.0065	0.0120 ± 0.0030
	FARVAT-XD	XCI	0.0406 ± 0.0055	0.0072 ± 0.0023
		E-XCI	0.0434 ± 0.0056	0.0092 ± 0.0026
	PedGene-Burden	XCI	0.0424 ± 0.0056	0.0078 ± 0.0024
		E-XCI	0.1541 ± 0.0100	0.0496 ± 0.0060
	PedGene-Kernel	XCI	0.0426 ± 0.0056	0.0078 ± 0.0024
		E-XCI	0.0638 ± 0.0068	0.0144 ± 0.0033
(A-2)	***FARVAT-XB***	XCI	0.0520 ± 0.0062	0.0114 ± 0.0029
		E-XCI	0.0512 ± 0.0061	0.0076 ± 0.0024
	FARVAT-XC	XCI	0.0506 ± 0.0061	0.0106 ± 0.0028
		E-XCI	0.0502 ± 0.0061	0.0100 ± 0.0028
	FARVAT-XO	XCI	0.0572 ± 0.0064	0.0125 ± 0.0031
		E-XCI	0.0513 ± 0.0061	0.0064 ± 0.0022
	FARVAT-XD	XCI	0.0432 ± 0.0056	0.0098 ± 0.0027
		E-XCI	0.0430 ± 0.0056	0.0086 ± 0.0026
	PedGene-Burden	XCI	0.0496 ± 0.0060	0.0092 ± 0.0026
		E-XCI	0.0736 ± 0.0072	0.0138 ± 0.0032
	PedGene-Kernel	XCI	0.0442 ± 0.0057	0.0076 ± 0.0024
		E-XCI	0.0518 ± 0.0061	0.0086 ± 0.0026
(A-3)	***FARVAT-XB***	XCI	0.0500 ± 0.0060	0.0104 ± 0.0028
		E-XCI	0.0500 ± 0.0060	0.0122 ± 0.0030
	FARVAT-XC	XCI	0.0500 ± 0.0060	0.0090 ± 0.0026
		E-XCI	0.0490 ± 0.0060	0.0106 ± 0.0028

	FARVAT-XO	XCI	0.0506 ± 0.0061	0.0116 ± 0.0030
		E-XCI	0.0465 ± 0.0058	0.0116 ± 0.0030
	FARVAT-XD	XCI	0.0454 ± 0.0058	0.0096 ± 0.0027
		E-XCI	0.0458 ± 0.0058	0.0082 ± 0.0025
	PedGene-Burden	XCI	0.0488 ± 0.0060	0.0082 ± 0.0025
		E-XCI	0.0488 ± 0.0060	0.0090 ± 0.0026
	PedGene-Kernel	XCI	0.0468 ± 0.0059	0.0088 ± 0.0026
		E-XCI	0.0422 ± 0.0056	0.0080 ± 0.0025
(A-4)	***FARVAT-XB***	XCI	0.0542 ± 0.0063	0.0132 ± 0.0032
		E-XCI	0.0506 ± 0.0061	0.0110 ± 0.0029
	FARVAT-XC	XCI	0.0444 ± 0.0057	0.0072 ± 0.0023
		E-XCI	0.0474 ± 0.0059	0.0114 ± 0.0029
	FARVAT-XO	XCI	0.0474 ± 0.0059	0.0100 ± 0.0028
		E-XCI	0.0556 ± 0.0063	0.0087 ± 0.0026
	FARVAT-XD	XCI	0.0428 ± 0.0056	0.0086 ± 0.0026
		E-XCI	0.0490 ± 0.0060	0.0086 ± 0.0026
	PedGene-Burden	XCI	0.0476 ± 0.0059	0.0118 ± 0.0030
		E-XCI	0.0878 ± 0.0078	0.0236 ± 0.0042
	PedGene-Kernel	XCI	0.0424 ± 0.0056	0.0084 ± 0.0025
		E-XCI	0.0494 ± 0.0060	0.0086 ± 0.0026
(A-5)	***FARVAT-XB***	XCI	0.0526 ± 0.0062	0.0110 ± 0.0029
		E-XCI	0.0466 ± 0.0058	0.0090 ± 0.0026
	FARVAT-XC	XCI	0.0506 ± 0.0061	0.0088 ± 0.0026
		E-XCI	0.0524 ± 0.0062	0.0110 ± 0.0029
	FARVAT-XO	XCI	0.0502 ± 0.0061	0.0093 ± 0.0027
		E-XCI	0.0518 ± 0.0061	0.0095 ± 0.0027
	FARVAT-XD	XCI	0.0442 ± 0.0057	0.0088 ± 0.0026
		E-XCI	0.0482 ± 0.0059	0.0080 ± 0.0025
	PedGene-Burden	XCI	0.0484 ± 0.0059	0.0124 ± 0.0031
		E-XCI	0.1815 ± 0.0107	0.0650 ± 0.0068
	PedGene-Kernel	XCI	0.0456 ± 0.0058	0.0070 ± 0.0023
		E-XCI	0.0626 ± 0.0067	0.0114 ± 0.0029

Table 4.3 Empirical type-1 error estimates for S-XCI. Empirical type-1 errors were calculated for five different family structures (A-1) – (A-5). The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01 significance levels.

Type of family	Statistics	Biological Model	α=.05(95%CI)	α=.01(95%CI)
(A-1)	***FARVAT-XB***	S-XCI to normal allele	0.0482 ± 0.0059	0.0100 ± 0.0028
		S-XCI to deleterious allele	0.0530 ± 0.0062	0.0108 ± 0.0029
	FARVAT-XC	S-XCI to normal allele	0.0562 ± 0.0064	0.0120 ± 0.0030
		S-XCI to deleterious allele	0.0494 ± 0.0060	0.0110 ± 0.0029
	FARVAT-XO	S-XCI to normal allele	0.0579 ± 0.0065	0.0134 ± 0.0032
		S-XCI to deleterious allele	0.0482 ± 0.0059	0.0118 ± 0.0030
	FARVAT-XD	S-XCI to normal allele	0.0448 ± 0.0057	0.0084 ± 0.0025
		S-XCI to deleterious allele	0.0444 ± 0.0057	0.0106 ± 0.0028
(A-2)	***FARVAT-XB***	S-XCI to normal allele	0.0524 ± 0.0062	0.0094 ± 0.0027
		S-XCI to deleterious allele	0.0512 ± 0.0061	0.0112 ± 0.0029
	FARVAT-XC	S-XCI to normal allele	0.0458 ± 0.0058	0.0076 ± 0.0024
		S-XCI to deleterious allele	0.0484 ± 0.0059	0.0092 ± 0.0026
	FARVAT-XO	S-XCI to normal allele	0.0453 ± 0.0058	0.0072 ± 0.0023
		S-XCI to deleterious allele	0.0543 ± 0.0063	0.0116 ± 0.0030
	FARVAT-XD	S-XCI	0.0410 ± 0.0055	0.0092 ± 0.0026

		to normal allele		
		S-XCI to deleterious allele	0.0428 ± 0.0056	0.0076 ± 0.0024
(A-3)	***FARVAT-XB***	S-XCI to normal allele	0.0502 ± 0.0061	0.0096 ± 0.0027
		S-XCI to deleterious allele	0.0514 ± 0.0061	0.0108 ± 0.0029
	FARVAT-XC	S-XCI to normal allele	0.0450 ± 0.0057	0.0082 ± 0.0025
		S-XCI to deleterious allele	0.0508 ± 0.0061	0.0090 ± 0.0026
	FARVAT-XO	S-XCI to normal allele	0.0472 ± 0.0059	0.0089 ± 0.0026
		S-XCI to deleterious allele	0.0534 ± 0.0062	0.0096 ± 0.0027
	FARVAT-XD	S-XCI to normal allele	0.0388 ± 0.0054	0.0088 ± 0.0026
		S-XCI to deleterious allele	0.0454 ± 0.0058	0.0096 ± 0.0027
(A-4)	***FARVAT-XB***	S-XCI to normal allele	0.0484 ± 0.0059	0.0090 ± 0.0026
		S-XCI to deleterious allele	0.0512 ± 0.0061	0.0088 ± 0.0026
	FARVAT-XC	S-XCI to normal allele	0.0432 ± 0.0056	0.0076 ± 0.0024
		S-XCI to deleterious allele	0.0470 ± 0.0059	0.0106 ± 0.0028
	FARVAT-XO	S-XCI to normal allele	0.0459 ± 0.0058	0.0119 ± 0.0030
		S-XCI to deleterious allele	0.0506 ± 0.0061	0.0107 ± 0.0028
	FARVAT-XD	S-XCI to normal allele	0.0350 ± 0.0051	0.0092 ± 0.0026
		S-XCI to deleterious allele	0.0430 ± 0.0056	0.0088 ± 0.0026

(A-5)	***FARVAT-XB***	S-XCI to normal allele	0.0512 ± 0.0061	0.0088 ± 0.0026
		S-XCI to deleterious allele	0.0512 ± 0.0061	0.0124 ± 0.0031
	FARVAT-XC	S-XCI to normal allele	0.0442 ± 0.0057	0.0100 ± 0.0028
		S-XCI to deleterious allele	0.0518 ± 0.0061	0.0094 ± 0.0027
	FARVAT-XO	S-XCI to normal allele	0.0443 ± 0.0057	0.0100 ± 0.0028
		S-XCI to deleterious allele	0.0514 ± 0.0061	0.0131 ± 0.0032
	FARVAT-XD	S-XCI to normal allele	0.0406 ± 0.0055	0.0088 ± 0.0026
		S-XCI to deleterious allele	0.0452 ± 0.0058	0.0116 ± 0.0030

Table 4.4 Empirical type-1 error estimates for XCI or E-XCI when prevalence for males and females are different. The prevalences for male and female are assumed to be 0.36 and 0.12, respectively. Empirical type-1 errors were calculated for five different family structures (A-1) – (A-5). The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01 significance levels.

Type of family	Statistics	Biological Model	α=.05(95%CI)	α=.01(95%CI)
(A-1)	***FARVAT-XB***	XCI	0.0486 ± 0.0060	0.0102 ± 0.0028
		E-XCI	0.0526 ± 0.0062	0.0142 ± 0.0033
	FARVAT-XC	XCI	0.0528 ± 0.0062	0.0132 ± 0.0032
		E-XCI	0.0486 ± 0.0060	0.0098 ± 0.0027
	FARVAT-XO	XCI	0.0469 ± 0.0059	0.0131 ± 0.0032
		E-XCI	0.0552 ± 0.0063	0.0123 ± 0.0031
	FARVAT-XD	XCI	0.0456 ± 0.0058	0.0130 ± 0.0031
		E-XCI	0.0460 ± 0.0058	0.0128 ± 0.0031
	PedGene-Burden	XCI	0.0492 ± 0.0060	0.0096 ± 0.0027
		E-XCI	0.6941 ± 0.0128	0.4477 ± 0.0138
	PedGene-Kernel	XCI	0.0468 ± 0.0059	0.0074 ± 0.0024
		E-XCI	0.1750 ± 0.0105	0.0572 ± 0.0064
(A-2)	***FARVAT-XB***	XCI	0.0490 ± 0.0060	0.0092 ± 0.0026
		E-XCI	0.0478 ± 0.0059	0.0082 ± 0.0025
	FARVAT-XC	XCI	0.0510 ± 0.0061	0.0102 ± 0.0028
		E-XCI	0.0470 ± 0.0059	0.0104 ± 0.0028
	FARVAT-XO	XCI	0.0536 ± 0.0062	0.0097 ± 0.0027
		E-XCI	0.0503 ± 0.0061	0.0085 ± 0.0025
	FARVAT-XD	XCI	0.0464 ± 0.0058	0.0086 ± 0.0026
		E-XCI	0.0398 ± 0.0054	0.0096 ± 0.0027
	PedGene-Burden	XCI	0.0456 ± 0.0058	0.0082 ± 0.0025
		E-XCI	0.8560 ± 0.0097	0.6717 ± 0.0130
	PedGene-Kernel	XCI	0.0472 ± 0.0059	0.0084 ± 0.0025
		E-XCI	0.2601 ± 0.0122	0.0918 ± 0.0080
(A-3)	***FARVAT-XB***	XCI	0.0454 ± 0.0058	0.0080 ± 0.0025
		E-XCI	0.0546 ± 0.0063	0.0132 ± 0.0032
	FARVAT-XC	XCI	0.0482 ± 0.0059	0.0114 ± 0.0029
		E-XCI	0.0444 ± 0.0057	0.0068 ± 0.0023

	FARVAT-XO	XCI	0.0514 ± 0.0061	0.0083 ± 0.0025
		E-XCI	0.0488 ± 0.0060	0.0101 ± 0.0028
	FARVAT-XD	XCI	0.0396 ± 0.0054	0.0072 ± 0.0023
		E-XCI	0.0444 ± 0.0057	0.0086 ± 0.0026
	PedGene-Burden	XCI	0.0442 ± 0.0057	0.0062 ± 0.0022
		E-XCI	0.8866 ± 0.0088	0.7237 ± 0.0124
	PedGene-Kernel	XCI	0.0468 ± 0.0059	0.0076 ± 0.0024
		E-XCI	0.2985 ± 0.0127	0.1126 ± 0.0088
(A-4)	***FARVAT-XB***	XCI	0.0466 ± 0.0058	0.0084 ± 0.0025
		E-XCI	0.0516 ± 0.0061	0.0110 ± 0.0029
	FARVAT-XC	XCI	0.0434 ± 0.0056	0.0086 ± 0.0026
		E-XCI	0.0480 ± 0.0059	0.0092 ± 0.0026
	FARVAT-XO	XCI	0.0441 ± 0.0057	0.0080 ± 0.0025
		E-XCI	0.0501 ± 0.0060	0.0122 ± 0.0030
	FARVAT-XD	XCI	0.0392 ± 0.0054	0.0082 ± 0.0025
		E-XCI	0.0414 ± 0.0055	0.0090 ± 0.0026
	PedGene-Burden	XCI	0.0472 ± 0.0059	0.0090 ± 0.0026
		E-XCI	0.7055 ± 0.0126	0.4719 ± 0.0138
	PedGene-Kernel	XCI	0.0482 ± 0.0059	0.0068 ± 0.0023
		E-XCI	0.1818 ± 0.0107	0.0552 ± 0.0063
(A-5)	***FARVAT-XB***	XCI	0.0534 ± 0.0062	0.0116 ± 0.0030
		E-XCI	0.0464 ± 0.0058	0.0096 ± 0.0027
	FARVAT-XC	XCI	0.0506 ± 0.0061	0.0092 ± 0.0026
		E-XCI	0.0448 ± 0.0057	0.0104 ± 0.0028
	FARVAT-XO	XCI	0.0592 ± 0.0065	0.0098 ± 0.0027
		E-XCI	0.0429 ± 0.0056	0.0093 ± 0.0027
	FARVAT-XD	XCI	0.0456 ± 0.0058	0.0110 ± 0.0029
		E-XCI	0.0412 ± 0.0055	0.0078 ± 0.0024
	PedGene-Burden	XCI	0.0506 ± 0.0061	0.0092 ± 0.0026
		E-XCI	0.3485 ± 0.0132	0.1532 ± 0.0100
	PedGene-Kernel	XCI	0.0424 ± 0.0056	0.0076 ± 0.0024
		E-XCI	0.0810 ± 0.0076	0.0208 ± 0.0040

Table 4.5 Empirical type-1 error estimates for S-XCI when prevalence for males and females are different. The prevalences for male and female are assumed to be 0.36 and 0.12, respectively. Empirical type-1 errors were calculated for five different family structures (A-1) – (A-5). The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01 significance levels.

Type of family	Statistics	Biological Model	α=.05(95%CI)	α=.01(95%CI)
(A-1)	***FARVAT-XB***	S-XCI to normal allele	0.0548 ± 0.0063	0.0102 ± 0.0028
		S-XCI to deleterious allele	0.0506 ± 0.0061	0.0086 ± 0.0026
	FARVAT-XC	S-XCI to normal allele	0.0540 ± 0.0063	0.0116 ± 0.0030
		S-XCI to deleterious allele	0.0532 ± 0.0062	0.0122 ± 0.0030
	FARVAT-XO	S-XCI to normal allele	0.0560 ± 0.0064	0.0110 ± 0.0029
		S-XCI to deleterious allele	0.0477 ± 0.0059	0.0099 ± 0.0027
	FARVAT-XD	S-XCI to normal allele	0.0452 ± 0.0058	0.0106 ± 0.0028
		S-XCI to deleterious allele	0.0438 ± 0.0057	0.0110 ± 0.0029
(A-2)	***FARVAT-XB***	S-XCI to normal allele	0.0574 ± 0.0064	0.0132 ± 0.0032
		S-XCI to deleterious allele	0.0508 ± 0.0061	0.0094 ± 0.0027
	FARVAT-XC	S-XCI to normal allele	0.0514 ± 0.0061	0.0118 ± 0.0030
		S-XCI to deleterious allele	0.0502 ± 0.0061	0.0092 ± 0.0026
	FARVAT-XO	S-XCI to normal allele	0.0615 ± 0.0067	0.0162 ± 0.0035
		S-XCI to deleterious allele	0.0446 ± 0.0057	0.0104 ± 0.0028

	FARVAT-XD	S-XCI to normal allele	0.0518 ± 0.0061	0.0132 ± 0.0032
		S-XCI to deleterious allele	0.0456 ± 0.0058	0.0106 ± 0.0028
(A-3)	***FARVAT-XB***	S-XCI to normal allele	0.0490 ± 0.0060	0.0094 ± 0.0027
		S-XCI to deleterious allele	0.0536 ± 0.0062	0.0094 ± 0.0027
	FARVAT-XC	S-XCI to normal allele	0.0484 ± 0.0059	0.0086 ± 0.0026
		S-XCI to deleterious allele	0.0478 ± 0.0059	0.0080 ± 0.0025
	FARVAT-XO	S-XCI to normal allele	0.0441 ± 0.0057	0.0078 ± 0.0024
		S-XCI to deleterious allele	0.0511 ± 0.0061	0.0096 ± 0.0027
	FARVAT-XD	S-XCI to normal allele	0.0382 ± 0.0053	0.0070 ± 0.0023
		S-XCI to deleterious allele	0.0416 ± 0.0055	0.0098 ± 0.0027
(A-4)	***FARVAT-XB***	S-XCI to normal allele	0.0540 ± 0.0063	0.0104 ± 0.0028
		S-XCI to deleterious allele	0.0464 ± 0.0058	0.0094 ± 0.0027
	FARVAT-XC	S-XCI to normal allele	0.0450 ± 0.0057	0.0082 ± 0.0025
		S-XCI to deleterious allele	0.0472 ± 0.0059	0.0078 ± 0.0024
	FARVAT-XO	S-XCI to normal allele	0.0502 ± 0.0061	0.0099 ± 0.0027
		S-XCI to deleterious allele	0.0451 ± 0.0058	0.0075 ± 0.0024
	FARVAT-XD	S-XCI to normal allele	0.0426 ± 0.0056	0.0080 ± 0.0025
		S-XCI	0.0382 ± 0.0053	0.0074 ± 0.0024

		to deleterious allele		
(A-5)	***FARVAT-XB***	S-XCI to normal allele	0.0520 ± 0.0062	0.0096 ± 0.0027
		S-XCI to deleterious allele	0.0512 ± 0.0061	0.0104 ± 0.0028
	FARVAT-XC	S-XCI to normal allele	0.0424 ± 0.0056	0.0074 ± 0.0024
		S-XCI to deleterious allele	0.0470 ± 0.0059	0.0086 ± 0.0026
	FARVAT-XO	S-XCI to normal allele	0.0518 ± 0.0061	0.0072 ± 0.0023
		S-XCI to deleterious allele	0.0518 ± 0.0061	0.0135 ± 0.0032
	FARVAT-XD	S-XCI to normal allele	0.0390 ± 0.0054	0.0074 ± 0.0024
		S-XCI to deleterious allele	0.0466 ± 0.0058	0.0098 ± 0.0027

Table 4.6 Empirical type-1 error estimates for XCI or E-XCI when prevalence for males and females are different. The prevalences for male and female are assumed to be 0.12 and 0.36, respectively. Empirical type-1 errors were calculated for five different family structures (A-1) – (A-5). The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01 significance levels.

Type of family	Statistics	Biological Model	α=.05(95%CI)	α=.01(95%CI)
(A-1)	***FARVAT-XB***	XCI	0.0558 ± 0.0064	0.0094 ± 0.0027
		E-XCI	0.0560 ± 0.0064	0.0138 ± 0.0032
	FARVAT-XC	XCI	0.0518 ± 0.0061	0.0090 ± 0.0026
		E-XCI	0.0486 ± 0.0060	0.0086 ± 0.0026
	FARVAT-XO	XCI	0.0542 ± 0.0063	0.0092 ± 0.0027
		E-XCI	0.0501 ± 0.0060	0.0159 ± 0.0035
	FARVAT-XD	XCI	0.0426 ± 0.0056	0.0080 ± 0.0025
		E-XCI	0.0464 ± 0.0058	0.0100 ± 0.0028
	PedGene-Burden	XCI	0.0472 ± 0.0059	0.0106 ± 0.0028
		E-XCI	0.3075 ± 0.0128	0.1290 ± 0.0093
	PedGene-Kernel	XCI	0.0486 ± 0.0060	0.0090 ± 0.0026
		E-XCI	0.0806 ± 0.0075	0.0138 ± 0.0032
(A-2)	***FARVAT-XB***	XCI	0.0540 ± 0.0063	0.0106 ± 0.0028
		E-XCI	0.0500 ± 0.0060	0.0110 ± 0.0029
	FARVAT-XC	XCI	0.0576 ± 0.0065	0.0134 ± 0.0032
		E-XCI	0.0522 ± 0.0062	0.0108 ± 0.0029
	FARVAT-XO	XCI	0.0580 ± 0.0065	0.0140 ± 0.0033
		E-XCI	0.0517 ± 0.0061	0.0118 ± 0.0030
	FARVAT-XD	XCI	0.0500 ± 0.0060	0.0094 ± 0.0027
		E-XCI	0.0450 ± 0.0057	0.0090 ± 0.0026
	PedGene-Burden	XCI	0.0408 ± 0.0055	0.0076 ± 0.0024
		E-XCI	0.0646 ± 0.0068	0.0168 ± 0.0036
	PedGene-Kernel	XCI	0.0460 ± 0.0058	0.0078 ± 0.0024
		E-XCI	0.0458 ± 0.0058	0.0054 ± 0.0020
(A-3)	***FARVAT-XB***	XCI	0.0524 ± 0.0062	0.0112 ± 0.0029
		E-XCI	0.0514 ± 0.0061	0.0114 ± 0.0029
	FARVAT-XC	XCI	0.0474 ± 0.0059	0.0100 ± 0.0028
		E-XCI	0.0482 ± 0.0059	0.0094 ± 0.0027

	FARVAT-XO	XCI	0.0532 ± 0.0062	0.0146 ± 0.0033
		E-XCI	0.0523 ± 0.0062	0.0104 ± 0.0028
	FARVAT-XD	XCI	0.0424 ± 0.0056	0.0098 ± 0.0027
		E-XCI	0.0470 ± 0.0059	0.0090 ± 0.0026
	PedGene-Burden	XCI	0.0504 ± 0.0061	0.0102 ± 0.0028
		E-XCI	0.0482 ± 0.0059	0.0124 ± 0.0031
	PedGene-Kernel	XCI	0.0368 ± 0.0052	0.0062 ± 0.0022
		E-XCI	0.0462 ± 0.0058	0.0064 ± 0.0022
(A-4)	***FARVAT-XB***	XCI	0.0522 ± 0.0062	0.0084 ± 0.0025
		E-XCI	0.0508 ± 0.0061	0.0112 ± 0.0029
	FARVAT-XC	XCI	0.0502 ± 0.0061	0.0110 ± 0.0029
		E-XCI	0.0512 ± 0.0061	0.0108 ± 0.0029
	FARVAT-XO	XCI	0.0463 ± 0.0058	0.0126 ± 0.0031
		E-XCI	0.0541 ± 0.0063	0.0103 ± 0.0028
	FARVAT-XD	XCI	0.0462 ± 0.0058	0.0084 ± 0.0025
		E-XCI	0.0486 ± 0.0060	0.0090 ± 0.0026
	PedGene-Burden	XCI	0.0462 ± 0.0058	0.0084 ± 0.0025
		E-XCI	0.0750 ± 0.0073	0.0192 ± 0.0038
	PedGene-Kernel	XCI	0.0440 ± 0.0057	0.0086 ± 0.0026
		E-XCI	0.0478 ± 0.0059	0.0074 ± 0.0024
(A-5)	***FARVAT-XB***	XCI	0.0520 ± 0.0062	0.0114 ± 0.0029
		E-XCI	0.0566 ± 0.0064	0.0114 ± 0.0029
	FARVAT-XC	XCI	0.0480 ± 0.0059	0.0088 ± 0.0026
		E-XCI	0.0454 ± 0.0058	0.0102 ± 0.0028
	FARVAT-XO	XCI	0.0546 ± 0.0063	0.0088 ± 0.0026
		E-XCI	0.0520 ± 0.0062	0.0116 ± 0.0030
	FARVAT-XD	XCI	0.0424 ± 0.0056	0.0096 ± 0.0027
		E-XCI	0.0436 ± 0.0057	0.0094 ± 0.0027
	PedGene-Burden	XCI	0.0460 ± 0.0058	0.0074 ± 0.0024
		E-XCI	0.1842 ± 0.0107	0.0670 ± 0.0069
	PedGene-Kernel	XCI	0.0398 ± 0.0054	0.0074 ± 0.0024
		E-XCI	0.0608 ± 0.0066	0.0118 ± 0.0030

Table 4.7 Empirical type-1 error estimates for S-XCI when prevalence for males and females are different. The prevalences for male and female are assumed to be 0.12 and 0.36, respectively. Empirical type-1 errors were calculated for five different family structures (A-1) – (A-5). The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01 significance levels.

Type of family	Statistics	Biological Model	α=.05(95%CI)	α=.01(95%CI)
(A-1)	***FARVAT-XB***	S-XCI to normal allele	0.0524 ± 0.0062	0.0118 ± 0.0030
		S-XCI to deleterious allele	0.0476 ± 0.0059	0.0092 ± 0.0026
	FARVAT-XC	S-XCI to normal allele	0.0494 ± 0.0060	0.0108 ± 0.0029
		S-XCI to deleterious allele	0.0498 ± 0.0060	0.0098 ± 0.0027
	FARVAT-XO	S-XCI to normal allele	0.0492 ± 0.0060	0.0123 ± 0.0031
		S-XCI to deleterious allele	0.0453 ± 0.0058	0.0087 ± 0.0026
	FARVAT-XD	S-XCI to normal allele	0.0456 ± 0.0058	0.0094 ± 0.0027
		S-XCI to deleterious allele	0.0424 ± 0.0056	0.0062 ± 0.0022
(A-2)	***FARVAT-XB***	S-XCI to normal allele	0.0500 ± 0.0060	0.0088 ± 0.0026
		S-XCI to deleterious allele	0.0494 ± 0.0060	0.0104 ± 0.0028
	FARVAT-XC	S-XCI to normal allele	0.0466 ± 0.0058	0.0094 ± 0.0027
		S-XCI to deleterious allele	0.0508 ± 0.0061	0.0108 ± 0.0029
	FARVAT-XO	S-XCI to normal allele	0.0453 ± 0.0058	0.0090 ± 0.0026
		S-XCI to deleterious allele	0.0484 ± 0.0059	0.0142 ± 0.0033

	FARVAT-XD	S-XCI to normal allele	0.0428 ± 0.0056	0.0088 ± 0.0026
		S-XCI to deleterious allele	0.0436 ± 0.0057	0.0090 ± 0.0026
(A-3)	***FARVAT-XB***	S-XCI to normal allele	0.0508 ± 0.0061	0.0106 ± 0.0028
		S-XCI to deleterious allele	0.0486 ± 0.0060	0.0098 ± 0.0027
	FARVAT-XC	S-XCI to normal allele	0.0476 ± 0.0059	0.0082 ± 0.0025
		S-XCI to deleterious allele	0.0446 ± 0.0057	0.0102 ± 0.0028
	FARVAT-XO	S-XCI to normal allele	0.0456 ± 0.0058	0.0106 ± 0.0028
		S-XCI to deleterious allele	0.0447 ± 0.0057	0.0100 ± 0.0028
	FARVAT-XD	S-XCI to normal allele	0.0410 ± 0.0055	0.0088 ± 0.0026
		S-XCI to deleterious allele	0.0420 ± 0.0056	0.0082 ± 0.0025
(A-4)	***FARVAT-XB***	S-XCI to normal allele	0.0494 ± 0.0060	0.0070 ± 0.0023
		S-XCI to deleterious allele	0.0562 ± 0.0064	0.0120 ± 0.0030
	FARVAT-XC	S-XCI to normal allele	0.0436 ± 0.0057	0.0080 ± 0.0025
		S-XCI to deleterious allele	0.0476 ± 0.0059	0.0094 ± 0.0027
	FARVAT-XO	S-XCI to normal allele	0.0507 ± 0.0061	0.0093 ± 0.0027
		S-XCI to deleterious allele	0.0511 ± 0.0061	0.0063 ± 0.0022
	FARVAT-XD	S-XCI to normal allele	0.0400 ± 0.0054	0.0066 ± 0.0022
		S-XCI	0.0494 ± 0.0060	0.0098 ± 0.0027

		to deleterious allele		
(A-5)	***FARVAT-XB***	S-XCI to normal allele	0.0436 ± 0.0057	0.0084 ± 0.0025
		S-XCI to deleterious allele	0.0482 ± 0.0059	0.0092 ± 0.0026
	FARVAT-XC	S-XCI to normal allele	0.0412 ± 0.0055	0.0082 ± 0.0025
		S-XCI to deleterious allele	0.0502 ± 0.0061	0.0106 ± 0.0028
	FARVAT-XO	S-XCI to normal allele	0.0459 ± 0.0058	0.0084 ± 0.0025
		S-XCI to deleterious allele	0.0569 ± 0.0064	0.0116 ± 0.0030
	FARVAT-XD	S-XCI to normal allele	0.0364 ± 0.0052	0.0064 ± 0.0022
		S-XCI to deleterious allele	0.0446 ± 0.0057	0.0092 ± 0.0026

In order to evaluate statistical efficiency, we considered five different extended family structures (A-1) – (A-5), and calculated the empirical power estimates for each. We assumed that there are 30 rare variants in each gene and 20 of them are causal. The number of deleterious causal rare variants was assumed to be 10, 12, 16, or 20. We assumed that h_a^2 was 0.01 and empirical power values at the 0.05 significance level were estimated with 5,000 replicates. Figure 4.2, 4.3, and 4.4 show that ***FARVAT-XB*** was the most powerful statistic if all risk variants are deleterious. If half of rare causal variants were deleterious and the other rare causal variants were protective, ***FARVAT-XC*** was the most powerful statistic. ***FARVAT-XO*** and ***FARVAT-XD*** were not always most efficient, but differences of power estimates among ***FARVAT-XO***, ***FARVAT-XD*** and the most efficient statistic were always small. It should be noted that ***FARVAT-XD*** is robust against the choice of mis-specified *d*. Figure 4.5 shows that PedGene-Burden is the most efficient statistic under E-XCI if all rare causal variants were deleterious, but it should be noted that empirical type-1 errors from PedGene-Burden were violated.

Figure 4.2 Empirical power estimates for random XCI. Empirical powers were calculated for five different extended family structures (A-1) – (A-5). h_a^2 was assumed to be 0.01 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Rare causal variants can have either deleterious or protective effect on disease, and the number of causal rare variants with deleterious effect was assumed to be 10, 12, 16, or 20.

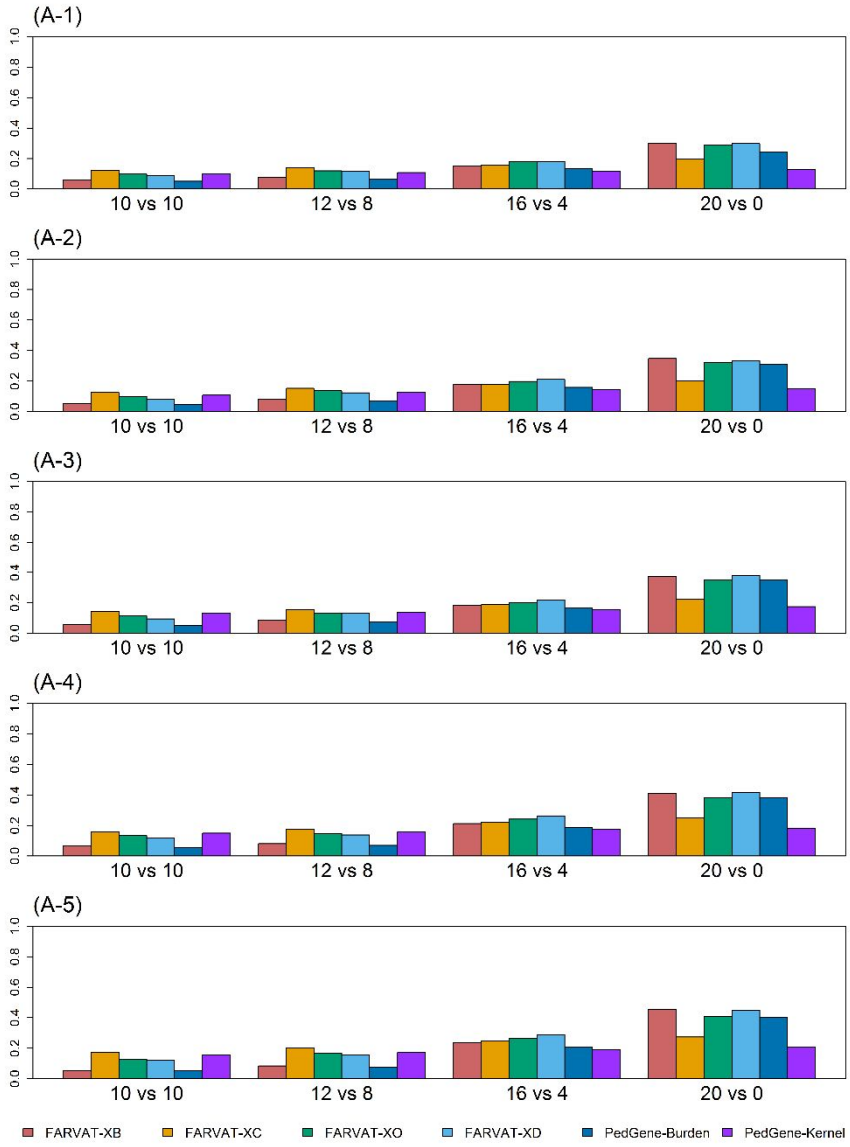


Figure 4.3 Empirical power estimates for S-XCI to normal allele. Empirical powers were calculated for five different extended family structures (A-1) – (A-5). h_a^2 was assumed to be 0.01 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Rare causal variants can have either deleterious or protective effect on disease, and the number of causal rare variants with deleterious effect was assumed to be 10, 12, 16, or 20.

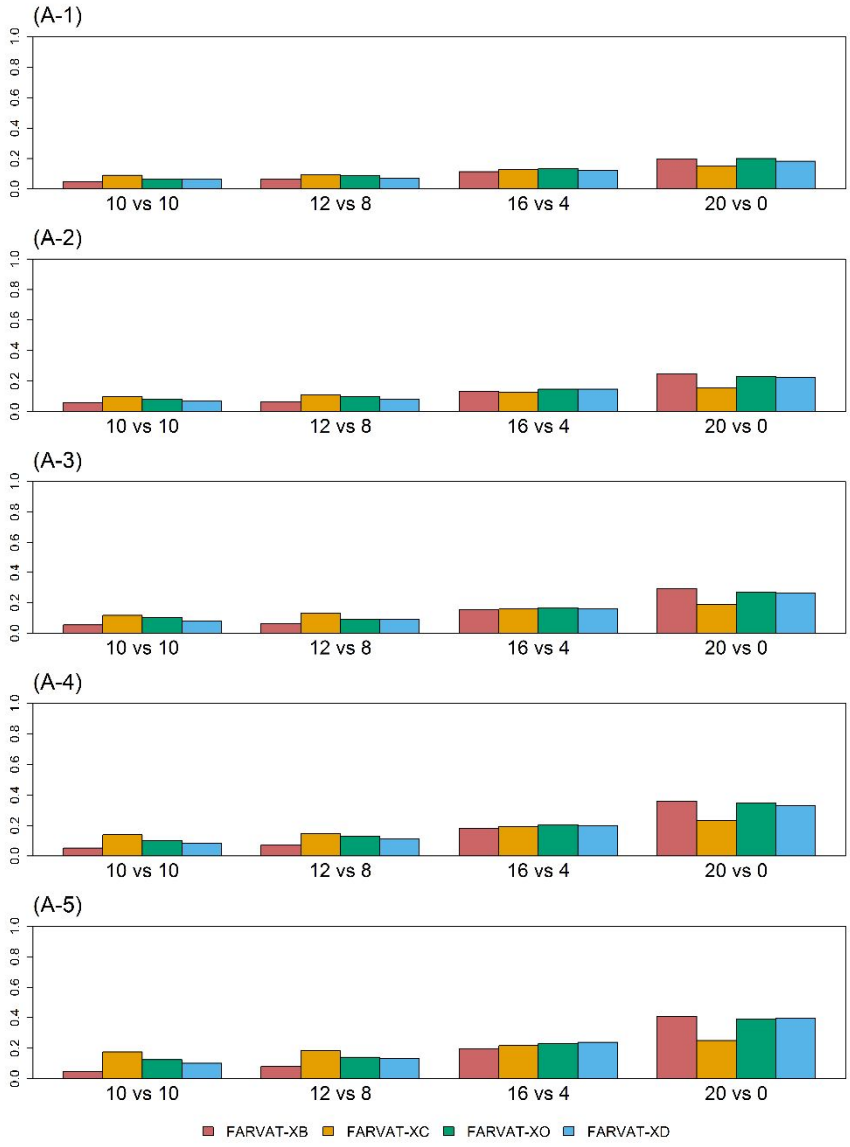


Figure 4.4 Empirical power estimates for S-XCI to deleterious allele. Empirical powers were calculated for five different extended family structures (A-1) – (A-5). h_a^2 was assumed to be 0.01 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Rare causal variants can have either deleterious or protective effect on disease, and the number of causal rare variants with deleterious effect was assumed to be 10, 12, 16, or 20.

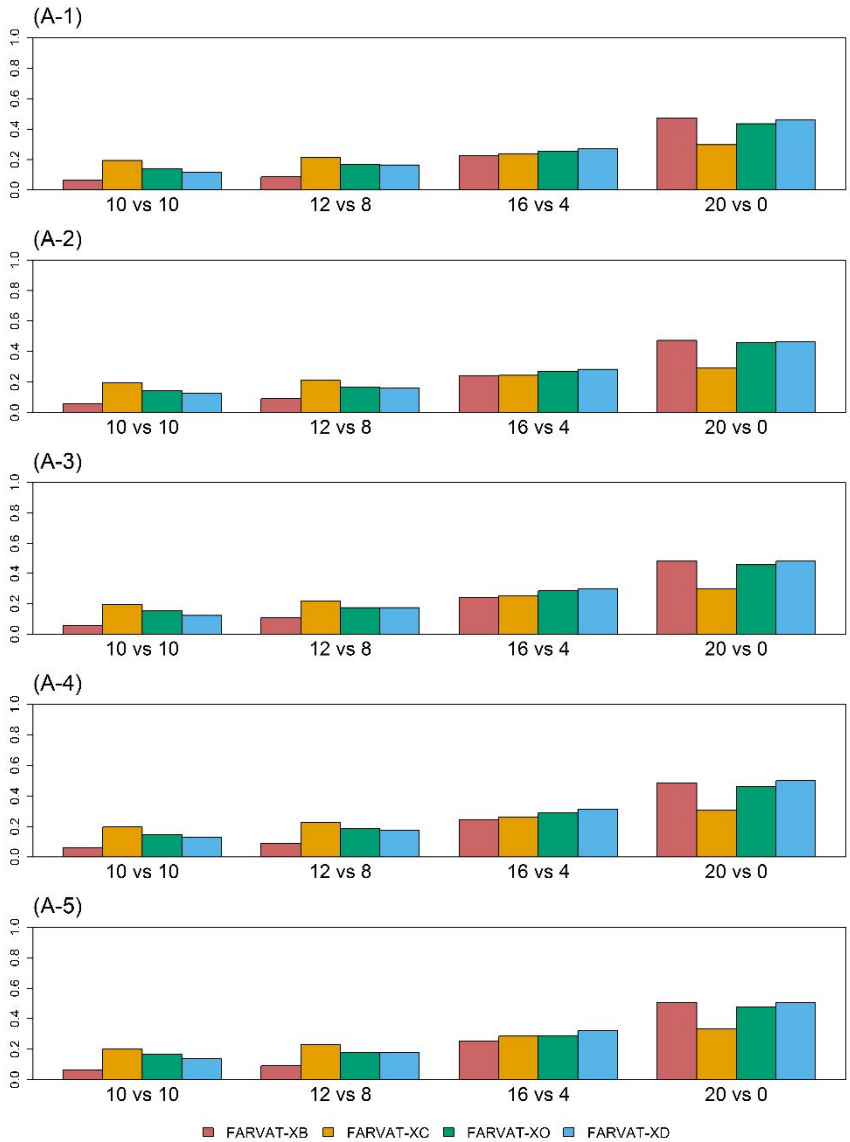
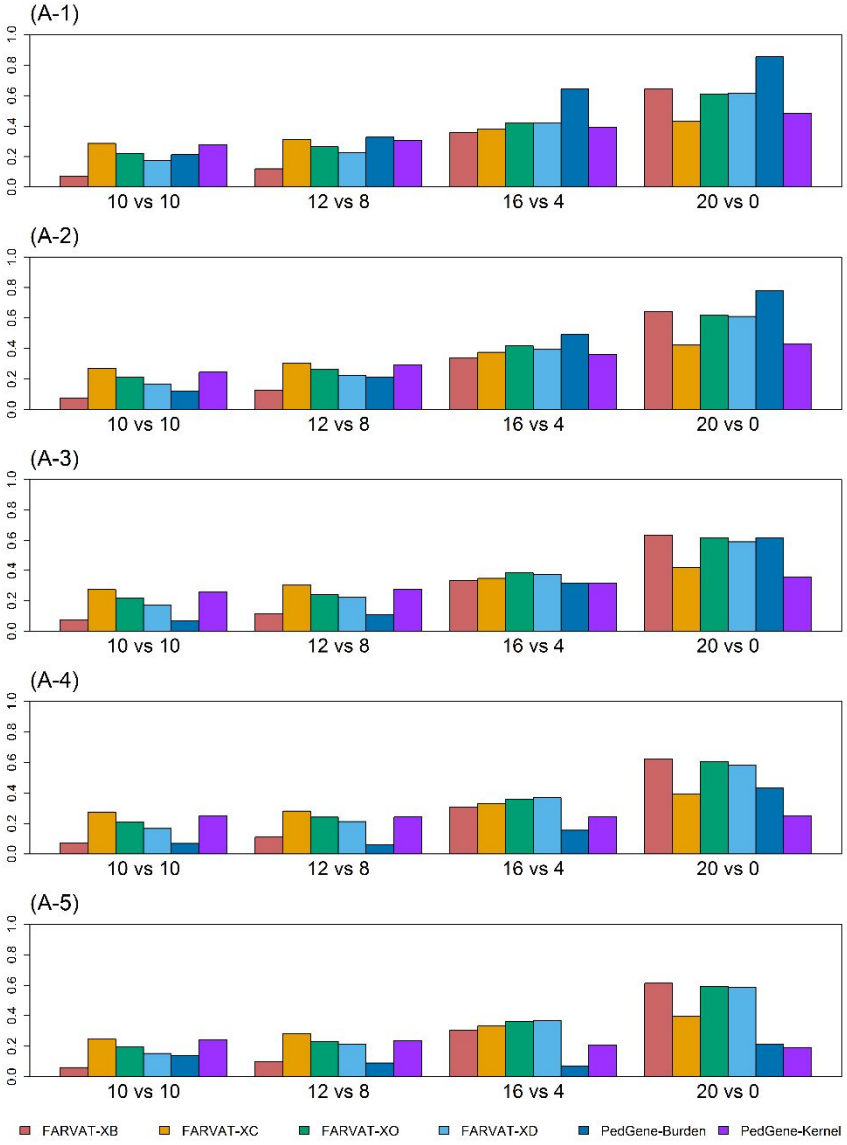


Figure 4.5 Empirical power estimates for E-XCI. Empirical powers were calculated for five different extended family structures (A-1) – (A-5). h_a^2 was assumed to be 0.01 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Rare causal variants can have either deleterious or protective effect on disease, and the number of causal rare variants with deleterious effect was assumed to be 10, 12, 16, or 20.



4.3.3 Evaluation with simulated data in the presence of population substructure

We estimated the type-1 error rate and power for the proposed methods in the presence of population substructure, and compared them to the same statistics from PedGene-Burden and PedGene-Kernel. In our proposed method, the presence of population substructure can be handled by adjusting the phenotypes with an EIGENSTRAT-based approach [Schaid, et al. 2013; Won, et al. 2012]. Specifically, principal component (PC) scores were estimated from the genetic relation matrix [Price, et al. 2006], and phenotypes were regressed on PC scores with the linear mixed model, which considers the correlation between family members. Residuals were then utilized as t_{ij} for the proposed methods. The type-1 error estimates for trios were calculated at the 0.05 and 0.01 significance levels with 5,000 replicates. We assumed that there were 30 rare variants available in a gene and family structure (B-1) and (B-2) in Figure 4.1. Table 4.8 shows inflation of type-1 error estimates for all methods unless phenotypes are adjusted with PC scores, and, in particular, PedGene-Kernel has the largest bias of type-1 error estimates.

The statistical efficiency was also evaluated with 5,000 replicates at the 0.05 significance level in the presence of population substructure. We assumed that h_a^2 is 0.05, and that there are 30 rare variants in a gene. Twenty rare variants were assumed to be causal, and each causal variant can have either deleterious or protective effects on phenotypes. Figure 4.6 shows that ***FARVAT-XB*** was the most efficient when all rare causal variants are

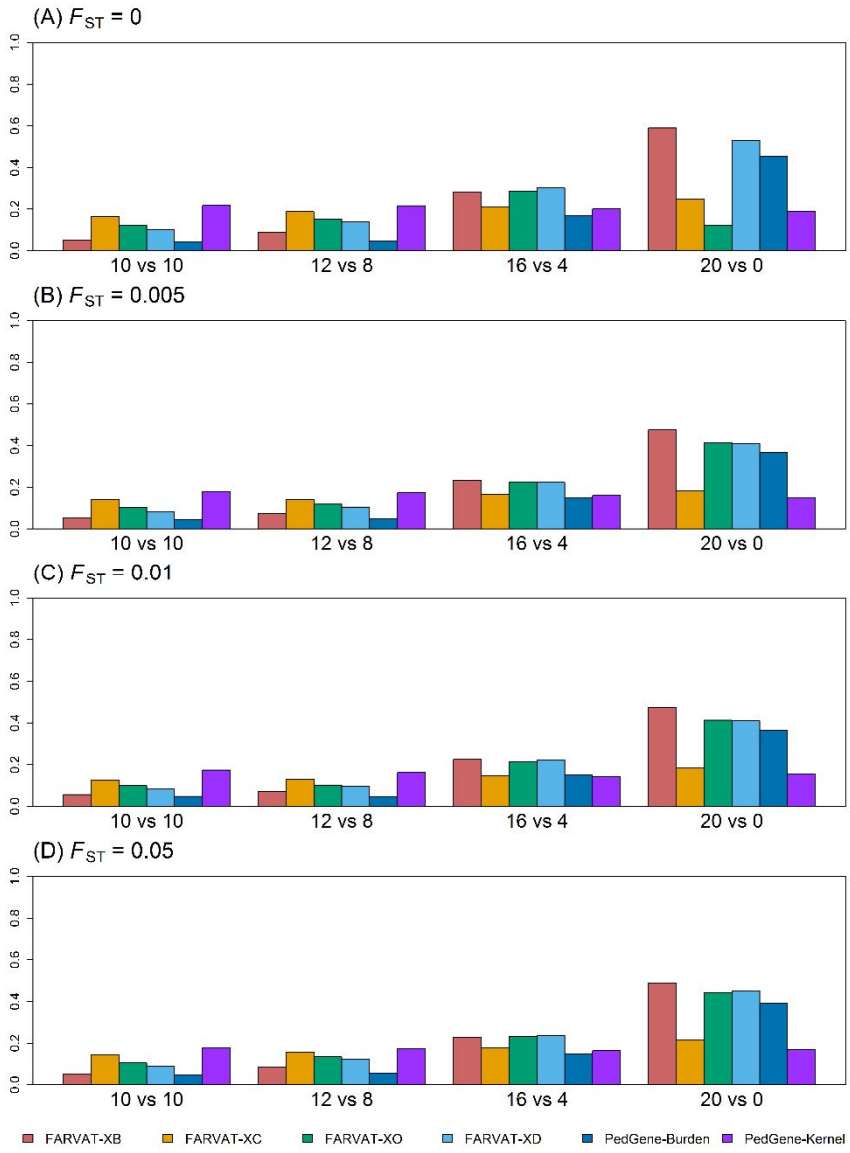
deleterious, and PedGene-Kernel was the most powerful if 50% of rare causal variants was deleterious. ***FARVAT-XO*** and ***FARVAT-XD*** are not always the most efficient, but their power loss when compared to the most efficient statistic is always small.

Table 4.8 Empirical type-1 error estimates for random XCI in the presence of population substructure. Empirical type-1 errors were calculated for two different trios structures (B-1) and (B-2). The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01 significance levels.

F_{ST}	Statistics	No adjustment with PC scores		Adjustment with PC scores	
		α=.05(95%CI)	α=.01(95%CI)	α=.05(95%CI)	α=.01(95%CI)
0	***FARVAT-XB***	0.0470 ± 0.0059	0.0100 ± 0.0028	0.0516 ± 0.0061	0.0086 ± 0.0026
	FARVAT-XC	0.0416 ± 0.0055	0.0072 ± 0.0023	0.0324 ± 0.0049	0.0044 ± 0.0018
	FARVAT-XO	0.0479 ± 0.0059	0.0111 ± 0.0029	0.0479 ± 0.0059	0.0087 ± 0.0026
	FARVAT-XD	0.0388 ± 0.0054	0.0070 ± 0.0023	0.348 ± 0.0051	0.0050 ± 0.0020
	PedGene-Burden	0.0380 ± 0.0053	0.0066 ± 0.0022	0.0386 ± 0.0053	0.0062 ± 0.0022
	PedGene-Kernel	0.0484 ± 0.0059	0.0070 ± 0.0023	0.0482 ± 0.0059	0.0068 ± 0.0023
0.005	***FARVAT-XB***	0.0458 ± 0.0058	0.0092 ± 0.0026	0.0440 ± 0.0057	0.0100 ± 0.0028
	FARVAT-XC	0.0436 ± 0.0057	0.0074 ± 0.0024	0.0324 ± 0.0049	0.0058 ± 0.0021
	FARVAT-XO	0.0489 ± 0.0060	0.0113 ± 0.0029	0.0388 ± 0.0054	0.0069 ± 0.0023
	FARVAT-XD	0.0344 ± 0.0051	0.0072 ± 0.0023	0.0296 ± 0.0047	0.0058 ± 0.0021
	PedGene-Burden	0.0400 ± 0.0054	0.0078 ± 0.0024	0.0366 ± 0.0052	0.0064 ± 0.0022
	PedGene-Kernel	0.0670 ± 0.0069	0.0090 ± 0.0026	0.0450 ± 0.0057	0.0082 ± 0.0025

0.01	***FARVAT-XB***	0.0544 ± 0.0063	0.0124 ± 0.0031	0.0530 ± 0.0062	0.0086 ± 0.0026
	FARVAT-XC	0.0538 ± 0.0063	0.0102 ± 0.0028	0.0308 ± 0.0048	0.0048 ± 0.0019
	FARVAT-XO	0.0522 ± 0.0062	0.0110 ± 0.0029	0.0429 ± 0.0056	0.0057 ± 0.0021
	FARVAT-XD	0.0440 ± 0.0057	0.0096 ± 0.0027	0.0346 ± 0.0051	0.0052 ± 0.0020
	PedGene-Burden	0.0482 ± 0.0059	0.0098 ± 0.0027	0.0434 ± 0.0056	0.0060 ± 0.0021
	PedGene-Kernel	0.0726 ± 0.0072	0.0150 ± 0.0034	0.0484 ± 0.0059	0.0066 ± 0.0022
0.05	***FARVAT-XB***	0.0564 ± 0.0064	0.0130 ± 0.0031	0.0424 ± 0.0056	0.0082 ± 0.0025
	FARVAT-XC	0.0818 ± 0.0076	0.0166 ± 0.0035	0.0410 ± 0.0055	0.0074 ± 0.0024
	FARVAT-XO	0.0646 ± 0.0068	0.0120 ± 0.0030	0.0355 ± 0.0051	0.0060 ± 0.0021
	FARVAT-XD	0.0590 ± 0.0065	0.0128 ± 0.0031	0.0356 ± 0.0051	0.0060 ± 0.0021
	PedGene-Burden	0.0544 ± 0.0063	0.0106 ± 0.0028	0.0340 ± 0.0050	0.0064 ± 0.0022
	PedGene-Kernel	0.1444 ± 0.0097	0.0426 ± 0.0056	0.0496 ± 0.0060	0.0082 ± 0.0025

Figure 4.6 Empirical power estimates for random XCI in the presence of population substructure. Empirical powers were calculated for two different trio structures (B-1) and (B-2). h_a^2 was assumed to be 0.05 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Rare causal variants can have either deleterious or protective effect on disease, and the number of causal rare variants with deleterious effect was assumed to be 10, 12, 16, or 20.



4.3.4 Evaluation of robustness against biological expression process

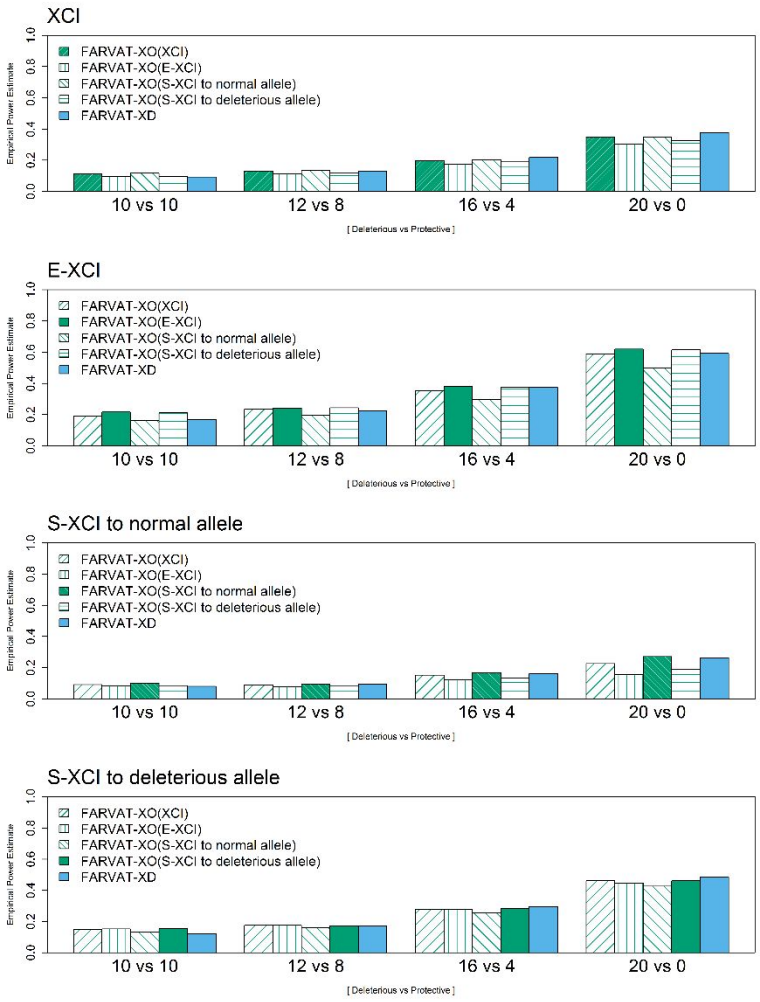
The gene expression process of X-linked variants is usually unknown, and the misspecified gene expression process may affect the performance of the proposed methods. We evaluated the robustness of the proposed methods with simulated data for (A-3) family structure. The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01 significance levels and Table 4.9 shows that type-1 error estimates of ***FARVAT-XO*** and ***FARVAT-XD*** consistently preserved the nominal significance levels. For evaluation of statistical powers, h_a^2 was assumed to be 0.01 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Figure 4.7 shows that ***FARVAT-XO*** with correctly specified biological model is the most efficient, but if it is misspecified, the power loss is usually substantial. ***FARVAT-XD*** is not the most efficient but the difference of its statistical powers with those for ***FARVAT-XO*** with correctly specified biological model is very small. Therefore, we can conclude that the performance of ***FARVAT-XO*** is affected by choice of d, and ***FARVAT-XD*** is generally a robust choice for various biological processes.

Table 4.9 Empirical type-1 error estimates when the gene expression process of X-linked variants are misspecified. Empirical type-1 errors were calculated for (A-3) family structure. The empirical type-1 error estimates were calculated with 5,000 replicates at the 0.05 and 0.01 significance levels.

Biological Model	Statistics	Assumed Biological Model	α=.05(95%CI)	α=.01(95%CI)
XCI	***FARVAT-XO***	XCI	0.0506 ± 0.0061	0.0116 ± 0.0030
		E-XCI	0.0535 ± 0.0062	0.0100 ± 0.0028
		S-XCI to normal allele	0.0506 ± 0.0061	0.0104 ± 0.0028
		S-XCI to deleterious allele	0.0531 ± 0.0062	0.0100 ± 0.0028
	FARVAT-XD	-	0.0454 ± 0.0058	0.0096 ± 0.0027
E-XCI	***FARVAT-XO***	XCI	0.0477 ± 0.0059	0.0078 ± 0.0024
		E-XCI	0.0465 ± 0.0058	0.0116 ± 0.0030
		S-XCI to normal allele	0.0531 ± 0.0062	0.0081 ± 0.0025
		S-XCI to deleterious allele	0.0446 ± 0.0057	0.0097 ± 0.0027
	FARVAT-XD	-	0.0458 ± 0.0058	0.0082 ± 0.0025

S-XCI to normal allele	***FARVAT-XO***	XCI	0.0484 ± 0.0060	0.0093 ± 0.0027
		E-XCI	0.0521 ± 0.0062	0.0109 ± 0.0029
		S-XCI to normal allele	0.0464 ± 0.0058	0.0081 ± 0.0025
		S-XCI to deleterious allele	0.0464 ± 0.0058	0.0101 ± 0.0028
	FARVAT-XD	-	0.0388 ± 0.0054	0.0088 ± 0.0026
S-XCI to deleterious allele	***FARVAT-XO***	XCI	0.0509 ± 0.0061	0.0108 ± 0.0029
		E-XCI	0.0542 ± 0.0063	0.0092 ± 0.0026
		S-XCI to normal allele	0.0530 ± 0.0062	0.0113 ± 0.0029
		S-XCI to deleterious allele	0.0521 ± 0.0062	0.0092 ± 0.0026
	FARVAT-XD	-	0.0454 ± 0.0058	0.0096 ± 0.0027

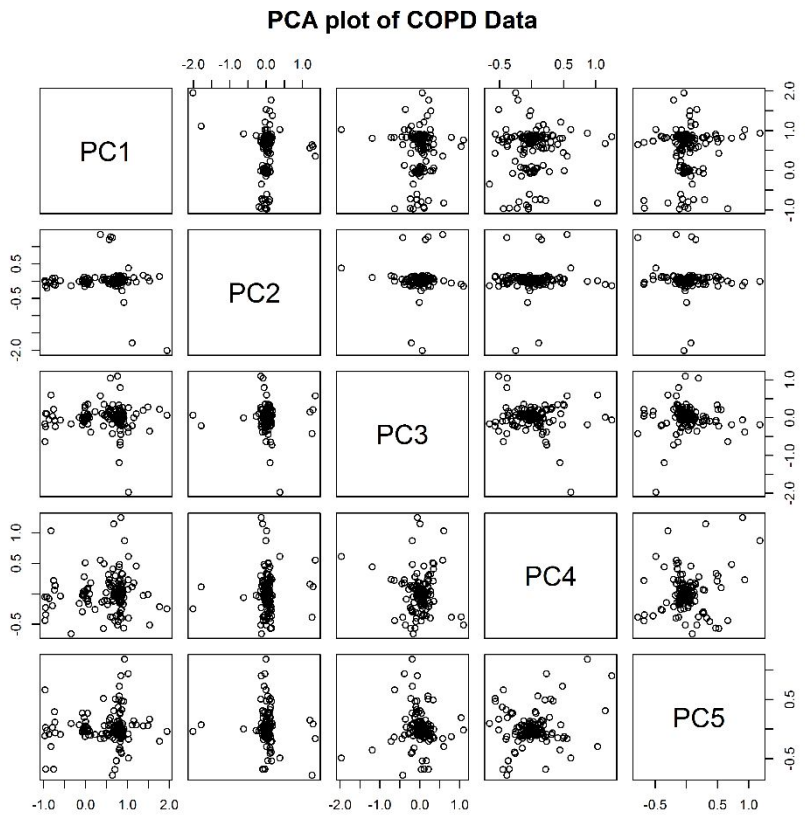
Figure 4.7 Empirical power estimates when the gene expression processes of X-linked variants are misspecified. Empirical powers were calculated for (A-3) family structure. h_a^2 was assumed to be 0.01 and the empirical power estimates were calculated with 5,000 replicates. We assumed that there are 30 rare variants, and among them 20 rare variants are causal. Rare causal variants can have either deleterious or protective effect on disease, and the number of causal rare variants with deleterious effect was assumed to be 10, 12, 16, or 20.



4.4 Application to chronic obstructive pulmonary disease data

The proposed methods were applied to rare variant association analyses of COPD using families from the Boston Early-Onset COPD Study with whole exome sequencing. Using moderate COPD or greater (FEV1 < 80% predicted with FEV1/FVC < 0.7) to define affection status, there were 64 unaffected males, 83 unaffected females, 55 affected males, and 100 affected females. There were 49 families and each family had at least two affected individuals. The whole exome of all individuals was sequenced with a Nimblegen V2 capture and Illumina platform. Sequencing data were preprocessed with the Genome Analysis ToolKit [McKenna, et al. 2010]. SNVs with Mendelian transmission errors, missing call rates (>1%), significant deviation from Hardy–Weinberg equilibrium ($P<10^{-8}$), read depth less than the average (12), and minor allele count of all variants in each gene (<5) were excluded. Seven genes in pseudo-autosomal regions and 186 genes with a single rare variant were excluded from our analyses. In total, we analyzed 629 rare variants in 183 genes on the X chromosome. There were 35,326 common autosomal variants with a MAF larger than 0.05, and they were utilized to calculate the genetic relationship matrix. Figure 4.8 shows the genetic relationships of the dataset on the first five PC scores.

Figure 4.8 Pairwise plots of PC scores.



Phenotypes were regressed with age, pack years, height, and 5 PC scores from the EIGENSTRAT method [Price, et al. 2006], and residuals were utilized as response variables to provide robustness of the proposed methods against population substructure. Figures 4.9 and 4.10 show quantile-quantile (QQ) plots of PedGene-Burden, PedGene-Kernel, and the proposed methods. QQ plots for PedGene-Burden and PedGene-Kernel show some evidence about inflation under random XCI and E-XCI, whereas the proposed methods are consistently valid. The most significant results were summarized in Table 4.10 The 0.05 exome-wide significant level adjusted by Bonferroni correction is 2.7E-04, and *q*-values [Storey 2002] were also provided in Table 4.10. Table 4.10 showed one exome-wide significant gene, *CXorf59* gene, with PedGene-Kernel for random XCI. However some inflation of results from PedGene-Kernel was confirmed with QQ plots and is not clear whether this significant association is valid. Some other promising results are also summarized in Table 4.10 and the second most significant results were obtained for the synovial sarcoma on X chromosome 5 (*SSX5*) gene using the proposed method. The significant association of *SYT-SSX* fusion gene with primary synovial sarcoma of the lung was reported [Hisaoka, et al. 1999], and the expression of SSX family genes (*SSX1*, *SSX2*, *SSX4*, and *SSX5*) were known to be related with lung cancer [Tureci, et al. 1998]. Furthermore, the *COL4A6* isoform have been shown to be more highly expressed in lung [Hudson, et al. 1993] and these significant results will be investigated as further studies.

Table 4.10 Most significant results from rare variant association analyses of COPD data.

Models	GENE	*M*	MAC		***FARVAT-XB***		***FARVAT-XC***		***FARVAT-XO***		***FARVAT-XD***		PedGene-Burden		PedGene-Kernel	
			Aff	Un.	*p*	*q*	*p*	*q*	*p*	*q*	*p*	*q*	*p*	*q*	*p*	*q*
XCI	*CXorf59*	2	2	7	0.165	0.798	0.016	0.590	0.026	0.425	0.039	0.552	0.001	0.205	1.6E-06	2.9E-04
	MTMR8	6	5	1	0.005	0.442	0.063	0.607	0.007	0.420	0.013	0.552	0.180	0.841	0.191	0.734
	SSX5	2	5	2	0.512	0.895	0.002	0.419	0.003	0.420	0.035	0.552	0.026	0.547	0.093	0.698
E-XCI	*CXorf59*	2	2	7	0.372	0.891	0.120	0.662	0.174	0.670	0.003	0.624	0.015	0.484	3.6E-05	0.006
	ELF4	4	13	5	0.919	0.957	0.934	0.955	0.964	0.969	0.784	0.903	0.003	0.484	0.034	0.330
	MTMR8	6	5	1	0.006	0.423	0.051	0.632	0.009	0.493	0.377	0.865	0.077	0.715	0.030	0.330
	SSX5	2	5	2	0.761	0.957	0.003	0.397	0.003	0.493	0.026	0.637	0.025	0.484	0.059	0.387
S-XCI to normal allele	*COL4A6*	7	18	21	0.002	0.316	0.022	0.677	0.003	0.459	0.007	0.552				
	CXorf59	2	2	7	0.094	0.783	0.005	0.524	0.008	0.459	0.039	0.552				
S-XCI to deleterious allele	*MTMR8*	6	5	1	0.005	0.482	0.050	0.579	0.007	0.424	0.013	0.552				
	SSX5	2	5	2	0.650	0.917	0.002	0.378	0.002	0.407	0.035	0.552				

Notes. The significant results for ***FARVAT-XB***, ***FARVAT-XC***, ***FARVAT-XO***, ***FARVAT-XD***, PedGene-Burden, and PedGene-Kernel are provided. The 0.05 exome-wide significant level adjusted by Bonferroni correction is 2.7E-04, and *q*-values [Storey 2002] are provided. *M* indicates the number of rare variant in a gene, and MAC indicates the minor allele counts.

Figure 4.9 QQ-plots of results from rare variant association analyses of COPD. QQ-plots are provided for PedGene-Burden, and PedGene-Kernel, and their 95% confidence interval is provided. Age, Pack-years of smoking, height, and 5 PCs were included as covariates for the linear mixed model and BLUP was utilized as offset.

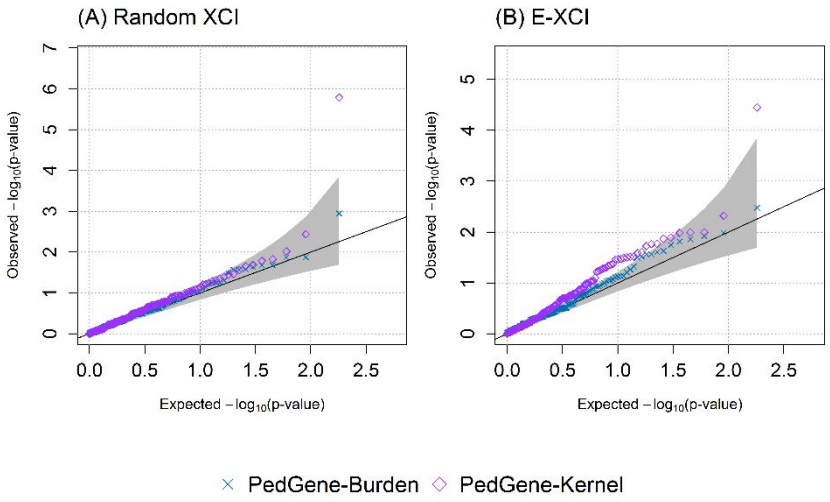
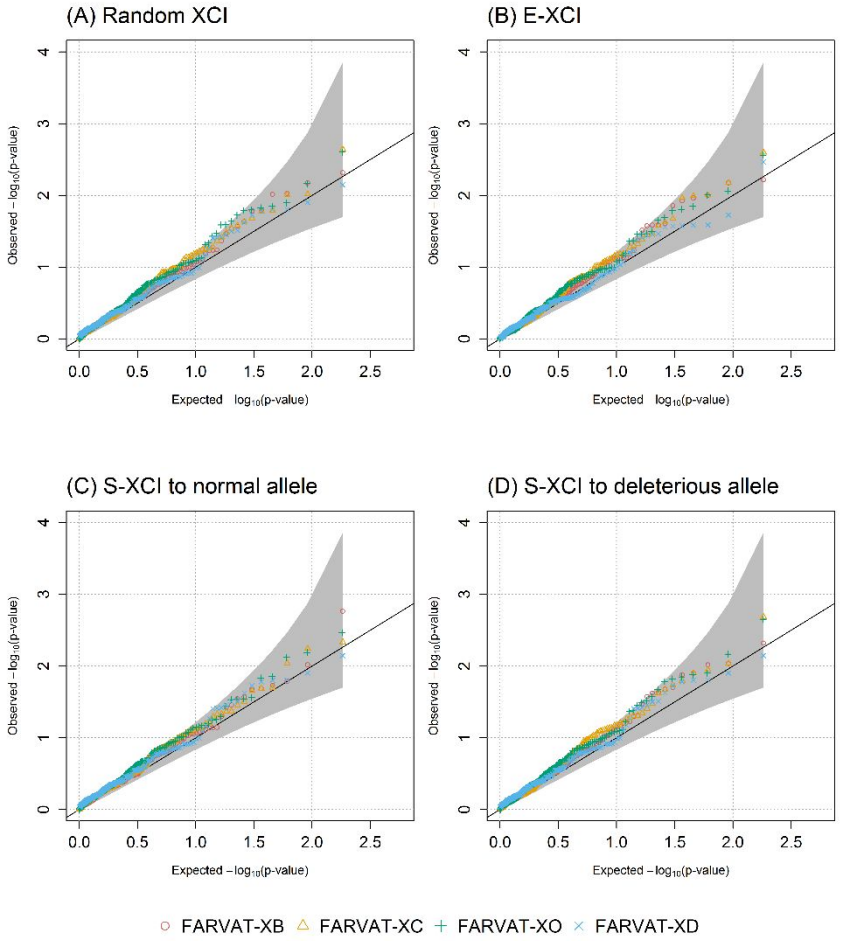


Figure 4.10 QQ plots of results from rare variant association analyses of COPD. QQ plots are provided for ***FARVAT-XB***, ***FARVAT-XC, FARVAT-XO***, and ***FARVAT-XD***, and their 95% confidence interval is provided. Age, Pack-years of smoking, height, and 5 PCs were included as covariates for the linear mixed model, and BLUP was utilized as offset.



4.5 Discussion

X-linked genes contribute to various biological mechanisms, including sexual dimorphism [Carrel and Willard 2005; Ober, et al. 2008; Tarpey, et al. 2009]. However, the complex biological processes associated with the expression of X-linked genes, such as random XCI, E-XCI, and S-XCI, complicate genetic association analyses with X-linked genes. Several methods for rare X-linked variants association analyses have been developed, but most cannot account for biologically plausible models. The limited discovery of significantly associated X-linked variants may be partially attributable to the absence of statistically efficient methods for detecting X-linked variants, and efficient analytical strategies for X-linked variants have been proposed as a potential mechanism to alleviate so-called "missing heritability" problems [Maher 2008; Manolio, et al. 2009].

In this chapter, we proposed a novel method for family-based association test of X-linked genes (***FARVATX***), which can accommodate random XCI, E-XCI, and S-XCI. The performance of ***FARVATX*** was evaluated with simulated data. We assumed that the magnitude of X-linked gene expression differed by gender and that the proportion of males and females in each family was different. The results from the simulation studies showed that PedGene-Burden and PedGene-Kernel statistics suffer from inflation of the type 1 error rate if the proportions of males and females are different or population substructure is present. However, ***FARVATX*** preserves the

nominal significance level in both the absence and presence of population substructure.

Furthermore, ***FARVATX*** is computationally less intensive than other available methods. Its application to sequencing data for COPD was completed within an hour. ***FARVATX*** software supports various input file formats, including plink and variant call format files, and multi-threaded analyses can be automatically conducted. The software for the proposed methods is written in C++ and can be downloaded from http://healthstat.snu.ac.kr/software/farvatx/.

Despite the analytical flexibility of the proposed methods, there are still some limitations. First, we found that the proposed methods are slightly conservative unless the sample size is sufficiently large, and it has been shown that small sample size adjustments by using resampling method leads to additional power improvement [Lee, et al. 2012a]. Second, the statistical power depends on the definition of rare variants, but it is still unclear. A variable threshold approach [Price, et al. 2010] that exhaustively searches the optimal MAF threshold may be a useful option for addressing this issue, and further extensions for the proposed methods are necessary. Third, the proposed methods assume that MAFs are same for males and females under the null hypothesis, and effects of each genetic variant for males and females are similar under the alternative hypothesis. If these are not satisfied, the false negative finding rates for the former and false positive findings rates for the

latter cannot be controlled, and males and females should be separately analyzed. These problems will be investigated in future studies.

The recent rapid improvement of sequencing technology provides the opportunity to identify rare X-linked variants associated with complex human diseases. However, our understanding of sex-specific genetic architecture and the biological processes associated with the expression of X-linked genes is still limited, and statistical methodology development to uncover them is necessary. The proposed methods may help us identify additional rare X-linked variants associated with complex traits, thereby leading to about a better understanding of the underlying biological processes associated with X-linked genes.

Chapter 5

Summary & Conclusions

Traditional genome-wide association studies (GWAS) have successfully identified more than 10,000 common variants associated with human complex traits in the last decade. Owing to the recent developments in next-generation sequencing (NGS) technologies, the NGS data will enable association studies on rare variants. Recently, many statistical methods have been proposed to detect novel rare variants related to human diseases. These methods, however, have been limited to the population-based design, and there are few statistical methods for analysis family data. In this thesis, we focused on family-based association test statistics for rare variants.

In chapter 3, we proposed a family-based rare-variant association test (***FARVAT***). The ***FARVAT*** method may be extended to various kinds of statistical methods such as the burden-type, variance component-type, or optimal-type methods depending on the choice of the working matrix in the

family-based design. The performance of ***FARVAT*** method was evaluated with simulated data in the presence or absence of the population substructure. The results of the simulation studies showed that $\boldsymbol{FARVAT_O}$ was generally an efficient choice for various disease models. The ***FARVAT*** method was applied to schizophrenia data of 36 trios from Germany. We found a genome-wide significant gene with $\boldsymbol{FARVAT_O}$. Furthermore, the ***FARVAT*** software was designed to be computationally efficient to analyze large NGS datasets.

In chapter 4, we focused on identifying rare variants on the X chromosome associated with complex traits. We proposed a family-based rare-variant association test for X-linked genes (***FARVATX***), which can accommodate random XCI, E-XCI, and S-XCI. The ***FARVATX*** method is an extension of ***FARVAT*** from autosomes to the X chromosome. The results of simulation studies revealed that the ***FARVATX*** method ensures global robustness in the presence of biased gender ratios in family structures, and ***FARVAT-XO*** and ***FARVAT-XD*** were generally an efficient choice for various types of biological models. In the analysis of chronic obstructive pulmonary disease (COPD), we found the synovial sarcoma on X chromosome 5 (SSX5) gene by mean of ***FARVATX***; this gene is known to be associated with lung function. Furthermore, the ***FARVATX*** software was found to be computationally more optimal than existing software. For instance, the real data analysis of the COPD dataset was completed within an hour using the FARVATX software.

In summary, the rare variant analysis with family members can lead to identification of rare variants that are more causal. This is because family members are genetically more homogeneous than unrelated samples. Therefore, family-based rare-variant association analysis seems to be both an efficient and cost-effective strategy, and the development of statistical methods for family-based sequencing data is necessary. We developed family-based rare-variant tests for autosomal and X chromosome genes and demonstrated good performance of the proposed methods in simulation studies and real data analyses.

Bibliography

Abkowitz JL, Taboada M, Shelton GH, Catlin SN, Guttorp P, Kiklevich JV. 1998. An X chromosome gene regulates hematopoietic stem cell kinetics. Proc Natl Acad Sci U S A 95(7):3862-6.

Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S and others. 2007. Medical sequencing at the extremes of human body mass. Am J Hum Genet 80(4):779-91.

Almasy L, Dyer TD, Peralta JM, Kent JW, Jr., Charlesworth JC, Curran JE, Blangero J. 2011. Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc 5 Suppl 9:S2.

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. Science 322(5903):881-8.

Amos-Landgraf JM, Cottle A, Plenge RM, Friez M, Schwartz CE, Longshore J, Willard HF. 2006. X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. Am J Hum Genet 79(3):493-9.

Armitage P. 1955. Tests for Linear Trends in Proportions and Frequencies. Biometrics 11(3):375-386.

Auer PL, Teumer A, Schick U, O'Shaughnessy A, Lo KS, Chami N, Carlson C, de Denus S, Dube MP, Haessler J and others. 2014. Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat Genet 46(6):629-34.

Auer PL, Wang G, Leal SM. 2013. Testing for rare variant associations in the presence of missing data. Genet Epidemiol 37(6):529-38.

Balding DJ, Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96(1-2):3-12.

Belmont JW. 1996. Genetic control of X inactivation and processes leading to X-inactivation skewing. Am J Hum Genet 58(6):1101-8.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57(1):289-300.

Blakemore AI, Meyre D, Delplanque J, Vatin V, Lecoeur C, Marre M, Tichet J, Balkau B, Froguel P, Walley AJ. 2009. A rare variant in the visfatin gene (NAMPT/PBEF1) is associated with protection from obesity. Obesity (Silver Spring) 17(8):1549-53.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140(2):783-96.

Brown CJ, Greally JM. 2003. A stain upon the silence: genes escaping X inactivation. Trends Genet 19(8):432-8.

Brown MB. 1975. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. Biometrics 31(4):987-992.

Busque L, Mio R, Mattioli J, Brais E, Blais N, Lalonde Y, Maragh M, Gilliland DG. 1996. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. Blood 88(1):59-65.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434(7031):400-4.

Cavalli-Sforza LL, Piazza A. 1993. Human genomic diversity in Europe: a summary of recent research and prospects for the future. Eur J Hum Genet 1(1):3-18.

Chagnon P, Provost S, Belisle C, Bolduc V, Gingras M, Busque L. 2005. Age-associated skewing of X-inactivation ratios of blood cells in normal females: a candidate-gene analysis approach. Exp Hematol 33(10):1209-14.

Chen H, Meigs JB, Dupuis J. 2013. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 37(2):196-204.

Chen LS, Hsu L, Gamazon ER, Cox NJ, Nicolae DL. 2012. An exponential combination procedure for set-based association tests in sequencing studies. Am J Hum Genet 91(6):977-86.

Choi S, Lee S, Cichon S, Nothen MM, Lange C, Park T, Won S. 2014. FARVAT: a family-based rare variant association test. Bioinformatics 30(22):3197-205.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 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.

Clayton D. 2008. Testing for association on the X chromosome. Biostatistics 9(4):593-600.

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305(5685):869-72.

Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH. 2006. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A 103(6):1810-5.

Davies RB. 1980a. Algorithm AS 155: The Distribution of a Linear Combination of $\chi 2$ Random Variables. Journal of the Royal Statistical Society. Series C (Applied Statistics) 29(3):323-333.

Davies RB. 1980b. The distribution of a linear combination of chi square random variables. Journal of the Royal Statistical Society. Series C (Applied Statistics) 29(3):323-33.

De G, Yip WK, Ionita-Laza I, Laird N. 2013. Rare variant analysis for family-based design. PLoS One 8(1):e48495.

Derkach A, Lawless JF, Sun L. 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):110-21.
Freidlin B, Zheng G, Li Z, Gastwirth JL. 2002. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 53(3):146-52.
Gaukrodger N, Mayosi BM, Imrie H, Avery P, Baker M, Connell JM, Watkins H, Farrall M, Keavney B. 2005. A rare variant of the leptin gene has large effects on blood pressure and carotid intima-medial thickness: a study of 1428 individuals in 248 families. J Med Genet 42(6):474-8.
Gibson G. 2011. Rare and common variants: twenty arguments. Nat Rev Genet 13(2):135-45.
Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82(1):100-12.
He Z, O'Roak BJ, Smith JD, Wang G, Hooker S, Santos-Cortez RL, Li B, Kan M, Krumm N, Nickerson DA and others. 2014. Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. Am J Hum Genet 94(1):33-46.
Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6(2):95-108.
Hisaoka M, Hashimoto H, Iwamasa T, Ishikawa K, Aoki T. 1999. Primary synovial sarcoma of the lung: report of two cases confirmed by molecular detection of SYT-SSX fusion gene transcripts. Histopathology 34(3):205-10.
Hu H, Roach JC, Coon H, Guthery SL, Voelkerding KV, Margraf RL, Durtschi JD, Tavtigian SV, Shankaracharya, Wu W and others. 2014. A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat Biotechnol 32(7):663-9.
Hudson BG, Reeders ST, Tryggvason K. 1993. Type IV collagen: structure, gene organization, and role in human diseases. Molecular basis of Goodpasture and Alport syndromes and diffuse leiomyomatosis. J Biol Chem 268(35):26033-6.
International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P and others. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851-61.
Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. 2011. A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet 7(2):e1001289.
Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. 2008. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40(5):592-9.
Jiang D, McPeek MS. 2014. Robust rare variant association testing for quantitative traits in samples with related individuals. Genet Epidemiol 38(1):10-20.

Knudsen GP, Pedersen J, Klingenberg O, Lygren I, Orstavik KH. 2007. Increased skewing of X chromosome inactivation with age in both blood and buccal cells. Cytogenet Genome Res 116(1-2):24-8.

Laird NM, Lange C. 2006. Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7(5):385-94.

Lander ES. 1996. The new genomics: global views of biology. Science 274(5287):536-9.

Lange C, Laird NM. 2002. Power calculations for a general class of family-based association tests: dichotomous traits. Am J Hum Genet 71(3):575-84.

Lee S, Abecasis GR, Boehnke M, Lin X. 2014. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet 95(1):5-23.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Team NGESP-ELP, Christiani DC, Wurfel MM, Lin X. 2012a. 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):224-37.

Lee S, Wu MC, Lin X. 2012b. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762-75.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83(3):311-21.

Liu DJ, Leal SM. 2010. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet 6(10):e1001156.

Liu H, Tang YQ, Zhang HH. 2009. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Computational Statistics & Data Analysis 53(4):853-856.

Lyon MF. 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature 190:372-3.

Ma C, Boehnke M, Lee S, Go TDI. 2015. Evaluating the Calibration and Power of Three Gene-Based Association Tests of Rare Variants for the X Chromosome. Genet Epidemiol 39(7):499-508.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2):e1000384.

Maher B. 2008. Personal genomes: The case of the missing heritability. Nature 456(7218):18-21.

Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118(5):1590-605.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-53.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9(5):356-69.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and others. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20(9):1297-1303.

McPeek MS, Wu X, Ober C. 2004. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60(2):359-67.

Minks J, Robinson WP, Brown CJ. 2008. A skewed view of X chromosome inactivation. J Clin Invest 118(1):20-3.

Morgenthaler S, Thilly WG. 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.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. PLoS Genet 7(3):e1001322.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324(5925):387-9.

Ober C, Loisel DA, Gilad Y. 2008. Sex-specific genetic architecture of human disease. Nat Rev Genet 9(12):911-22.

Ott J. 1989. Statistical properties of the haplotype relative risk. Genet Epidemiol 6(1):127-30.

Plenge RM, Stevenson RA, Lubs HA, Schwartz CE, Willard HF. 2002. Skewed X-chromosome inactivation is a common feature of X-linked mental retardation disorders. Am J Hum Genet 71(1):168-73.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86(6):832-8.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904-9.

Psychiatric GCCC, Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A and others. 2009. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. Am J Psychiatry 166(5):540-56.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273(5281):1516-7.

Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen JC. 2009. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest 119(1):70-9.

Sasieni PD. 1997. From genotypes to genes: doubling the sample size. Biometrics 53(4):1253-61.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11):1576-83.

Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. 2013. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol 37(5):409-18.

Sha Q, Zhang S. 2014. A rare variant association test based on combinations of single-variant tests. Genet Epidemiol 38(6):494-501.

Shapiro LJ, Mohandas T, Weiss R, Romeo G. 1979. Non-inactivation of an x-chromosome locus in man. Science 204(4398):1224-6.

Sharp A, Robinson D, Jacobs P. 2000. Age- and tissue-specific variation of X chromosome inactivation ratios in normal women. Hum Genet 107(4):343-9.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S and others. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445(7130):881-5.

Smit LA, Siroux V, Bouzigon E, Oryszczyn MP, Lathrop M, Demenais F, Kauffmann F, Epidemiological Study on the G, Environment of Asthma BH, Atopy Cooperative G. 2009. CD14 and toll-like receptor gene polymorphisms, country living, and asthma in adults. Am J Respir Crit Care Med 179(5):363-8.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52(3):506-16.

Storey JD. 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B-Statistical Methodology 64:479-498.

Sun J, Zheng Y, Hsu L. 2013. A unified mixed-effects model for rare-variant association in sequencing studies. Genet Epidemiol 37(4):334-44.

Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, O'Meara S, Latimer C, Dicks E, Menzies A and others. 2009. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat Genet 41(5):535-43.

Thornton T, McPeek MS. 2007. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet 81(2):321-37.

Thornton T, McPeek MS. 2010. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet 86(2):172-84.

Thornton T, Zhang Q, Cai X, Ober C, McPeek MS. 2012. XM: association testing on the X-chromosome in case-control samples with related individuals. Genet Epidemiol 36(5):438-50.

Tureci O, Chen YT, Sahin U, Gure AO, Zwick C, Villena C, Tsang S, Seitz G, Old LJ, Pfreundschuh M. 1998. Expression of SSX genes in human tumors. Int J Cancer 77(1):19-23.

Visscher PM. 2008. Sizing up human height variation. Nat Genet 40(5):489-90.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. Am J Hum Genet 90(1):7-24.

Wang J, Yu R, Shete S. 2014. X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. Genet Epidemiol 38(6):483-93.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L and others. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42(Database issue):D1001-6.

Won S, Elston RC. 2008. The power of independent types of genetic information to detect association in a case-control study design. Genet Epidemiol 32(8):731-56.

Won S, Lange C. 2013. A general framework for robust and efficient association analysis in family-based designs: quantitative and dichotomous phenotypes. Stat Med 32(25):4482-98.

Won S, Lu Q, Bertram L, Tanzi RE, Lange C. 2012. On the meta-analysis of genome-wide association studies: a robust and efficient approach to combine population and family-based studies. Hum Hered 73(1):35-46.

Won S, Wilk JB, Mathias RA, O'Donnell CJ, Silverman EK, Barnes K, O'Connor GT, Weiss ST, Lange C. 2009. On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. PLoS Genet 5(11):e1000741.

Wong CC, Caspi A, Williams B, Houts R, Craig IW, Mill J. 2011. A longitudinal twin study of skewed X chromosome-inactivation. PLoS One 6(3):e17873.

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

Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG. 2011. A probabilistic disease-gene finder for personal genomes. Genome Res 21(9):1529-42.

Zheng G, Joo J, Zhang C, Geller NL. 2007. Testing association for markers on the X chromosome. Genet Epidemiol 31(8):834-43.

Zhu Y, Xiong M. 2012. Family-Based Association Studies for Next-Generation Sequencing. American Journal of Human Genetics 90(6):1028-1045.

초 록

전장유전체연구(GWAS)는 100,000 – 1,000,000 개의 단일 염기변이(SNP)를 역학자료 혹은 임상자료와 연계시켜 특정 질환과 연관성이 있는 유전자나 원인-유전자 위치를 규명하는 통계 유전체 분석기법을 의미한다. 인간유전체사업(Human Genome Project)의 성공으로 밝혀진 SNP 에 대한 정보를 토대로 전장유전체 연구가 활발히 진행되었다. 그 결과 전장 유전체 연구를 통하여 복합질환(complex disease)과 관련된 후보 유전자들이 성공적으로 발굴되었다.

전장유전체연구가 성공적으로 진행되었지만, 현재까지 보고된 유전변이들이 질환의 유전율(heritability)을 설명하는 비율이 높지 않다는 문제가 제기되었다. 이러한 문제를 해결하기 위한 대안 중 하나로 희귀변이(rare variant) 연구의 중요성이 대두되었다.

희귀변이는 공통변이(common variant)와 달리, 소수의 사람들에게서만 발견된다는 점에서 개별인자 단위의 통계적 유의성을 얻기가 쉽지 않으므로 기존의 전장유전체 연관성 분석 방법을 적용하는데 어려움이 있다. 이러한 희귀변이를 유전자 단위로 병합하여 검정하는 방법들이 활발히 연구되고 있으며, 제안된 통계량들은 각기 장단점을 가지고 있고 여러 가지 가정 하에 일관된 검정력을 가지는 방법은 거의 없다. 또한 가족 데이터 기반의 희귀변이 연구를 위한 통계 분석 모형이 활발히 연구되지 않은 상황이다. 그 외에, 남녀에 따라 유전자 발현이 달라지는 X 염색체 상의 유전자에 대한 연관성 분석 방법론 개발이 미비한 실정이다.

이에 본 연구에서는 가족 데이터 기반의 희귀변이 연관분석을 위한 새로운 분석방법(*FARVAT*)과 이를 X chromosome 연구에

확장한 통계모형(*FARVATX*)을 제안하였다. *FARVAT* 분석모형은 quasi-likelihood 기반의 가족 데이터를 효율적으로 분석하는 방법으로, 가족 구성원들 간의 유전적 유사성은 kinship coefficient 정보를 활용하였다. *FARVAT* 모형은 burden type, variance component type, optimal type 통계량으로 확장이 가능하다. X 염색체 연구에서는, X 염색체불활성화(X chromosome inactivation, XCI)에 대한 다양한 생물학적 모델(XCI, escape-XCI, skewness-XCI)을 고려한 가족기반 rare variant X 염색체 분석모형(*FARVATX*)을 제안하였다. 새롭게 제안한 방법들은 다양한 시뮬레이션 연구를 수행하였고, 실제 데이터에 적용하여 특정질환과 관련된 후보 유전자들을 발굴하였다. 본 논문에서 제시된 방법들이 복합질환에 영향을 미치는 후보 유전자들을 효과적으로 발굴하고 질병의 발생 기작(mechanism)을 연구하는데 활용될 수 있을 것으로 기대된다.

주요어: 전장유전체연관성분석, 차세대시퀀싱자료분석, 희귀변이 연관성테스트, 가족 기반 디자인, X 염색체

학 번: 2011-30927