CORE
Metadata, citation and similar papers at core.ac.uk
Provided by D-Scholarship@Pitt

ASSOCIATION ANALYSIS BETWEEN BINARY TRAITS AND COMMON OR RARE GENETIC VARIANTS ON FAMILY-BASED DATA

by

Jia Jia

M.B.A., University of New Haven, 2010

B.S., Tianjin University of Commerce, China, 2008

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Jia Jia

It was defended on

2/25/2015

and approved by

Daniel E. Weeks, PhD

Professor

Department of Human Genetics

Graduate School of Public Health

University of Pittsburgh

Eleanor Feingold, PhD

Professor

Department of Human Genetics

Graduate School of Public Health

University of Pittsburgh

George C. Tseng, ScD

Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Wei Chen, PhD

Assistant Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Dissertation Director: **Daniel E. Weeks**, PhD

Professor

Department of Human Genetics

Graduate School of Public Health

University of Pittsburgh

Copyright © by Jia Jia
2015

ASSOCIATION ANALYSIS BETWEEN BINARY TRAITS AND COMMON OR RARE GENETIC VARIANTS ON FAMILY-BASED DATA

Jia Jia, PhD

University of Pittsburgh, 2015

ABSTRACT

Association studies test for genetic variation influencing disease risk. We explore here the application and development of statistics for binary traits on family data. There are two main areas of focus: the first on comparing existing single-variant tests, and the second on developing a gene-based test.

In the first part, we carried out a comparative study by applying 42 family-based association test statistics on different family-based datasets, which are simulated under a variety of scenarios (varying levels of linkage disequilibrium; dominant, additive, and recessive disease models; a variety of family structures). We have compared the Type I error, power and robustness of all the statistics. The results show that, when testing the null hypothesis of no association and no linkage, among the statistics that have well-behaved Type I error, the More powerful Quasi-likelihood Score test has the highest power and high robustness.

In the second part, motivated by a need for powerful gene-based association statistics on family-based data for binary traits, we have proposed a new test statistic, which is based on a mixed model framework, Laplace's method and a variance component score test. We have compared the Type I error rates and power of our new statistic and six existing statistics by simulating different scenarios (varying the number and effect size of risk and protective variants). Our proposed statistic shows well-behaved Type I error and high power in some scenarios.

The insights gathered here may improve public health by providing information on how to effectively utilize association methods to detect genetic variants that are related to disease. Ultimately, they should help improve the understanding of disease etiology.

TABLE OF CONTENTS

1.0 INTRODUCTION 1

1.1 Overview 1

1.2 General background 1

1.2.1 Trait and marker 1

1.2.2 Association analysis 2

1.2.3 Population and family data 2

1.2.4 Linkage and association 2

1.2.5 Common and rare variants 5

2.0 A SIMULATION-BASED COMPARATIVE STUDY OF FAMILY-BASED ASSOCIATION TESTS 6

2.1 Motivation 6

2.2 Approach 7

2.2.1 Null Hypothesis 7

2.2.2 Simulation Description 16

2.3 Results 20

2.3.1 Clusters 20

2.3.2 Type I error 22

2.3.3 Power 26

2.4 Discussion 29

2.4.1 Population stratification 31

2.4.2 Mis-specified family structure 33

2.4.3 Test for association in the presence of linkage 34

2.4.4 Ascertainment criteria and study design . . . 35
3.0 FAMILY-BASED RARE VARIANTS ASSOCIATION ANALYSIS FOR BINARY TRAITS . . . 37
3.1 Background and Motivation . . . 37
3.1.1 Existing methods . . . 38
3.1.1.1 Burden tests . . . 39
3.1.1.2 Bi-directional (Kernel) tests . . . 39
3.1.1.3 Combined and PCA tests . . . 41
3.1.2 Familial correlation . . . 41
3.1.3 Continuous traits versus binary traits . . . 42
3.2 Approach . . . 43
3.2.1 Proposed Method: Model setting . . . 44
3.2.2 Proposed Method: Inference method . . . 46
3.2.2.1 Quasi-Likelihood . . . 46
3.2.2.2 Score function . . . 51
3.2.2.3 Information matrix . . . 52
3.2.2.4 Q-statistic . . . 53
3.3 Simulation . . . 53
3.4 Results . . . 58
3.4.1 Type I error . . . 59
3.4.2 Power . . . 63
3.5 Discussion . . . 69
3.5.1 Weighting matrix . . . 70
3.5.2 Inflated Type I error . . . 71
3.5.3 Population stratification and family structure mis-specification . . . 73
3.5.4 Untyped subjects . . . 73
4.0 SUMMARY AND FUTURE WORK . . . 75
4.1 A simulation-based comparative study of family-based association tests . . . 75
4.2 Family-based rare variants association analysis for binary traits . . . 75
APPENDIX A. SUPPLEMENTARY TABLES AND FIGURES . . . 79

APPENDIX B. SIMULATED FAMILIES FOR RARE VARIANT ASSOCIATION ANALYSIS 92
APPENDIX C. R IMPLEMENTATION OF THE QTEST 118
BIBLIOGRAPHY 121

LIST OF TABLES

2.1 Abbreviations, null Hypothesis and descriptions of all statistics evaluated in this study 8

2.2 Simulation Scenarios with Marker/Disease allele frequency = 0.2/0.2 16

2.3 Counts of Type I error behavior and power across 15 different scenarios. . . . 23

3.1 Example of Burden test collapsing genotype 39

3.2 Eight simulated scenarios 59

A1 Type I error and power for all statistics across all scenarios. 79

LIST OF FIGURES

1.1 Example of linkage and association in family-based data. 4

2.1 Examples of simulated family structures. The numbers of offspring are randomly generated from a negative binomial distribution. 18

2.2 Hierarchical clustering plot based on Manhattan distance of p-values under Null NL across fully typed family structures and penetrance models. 21

2.3 Power of selected statistics with well-behaved Type I error at 0.05 alpha level, with D'= 0.6, across all simulated scenarios. The bars within each statistic are ordered according to the legends. g_tdt does not work on 2genUP families. Error bars are 95% confidence intervals calculated based on 200 replicates. . . 27

3.1 An example of simulated family structure. The number of offspring are randomly generated from a negative binomial distribution. Please see Appendix B for all the 25 simulated families. 54

3.2 Type I error at a gene independent of the trait locus under eight trait simulation scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.05 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. The boundaries of the 95% Confidence Interval are marked out with two black lines. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights. . . 60

3.3 Type I error at a gene independent of the trait locus under eight trait simulation scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.01 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. The boundaries of the 95% Confidence Interval are marked out with two black lines. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights. . . 61

3.4 Type I error at a gene independent of the trait locus under eight trait simulation scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.001 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. The boundaries of the 95% Confidence Interval are marked out with two black lines. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights. . . 62

3.5 Power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.05 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights. . . 65

3.6 Power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.01 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights. . . 66

3.7 Power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.001 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights. . . 67

A1 Hierarchical clustering plot based on Euclidean distance of p-values under Null NL across fully typed family structures and penetrance models 88

A2 Adjusted power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.05 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights. 89

A3 Adjusted power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.01 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights. 90

A4 Adjusted power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.001 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^{+} = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^{-} = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights. 91

1.0 INTRODUCTION

1.1 OVERVIEW

The unifying aim of this dissertation is to research family-based association test statistics for binary traits and for both common and rare variants. In Chapter 2, we conducted a simulation-based comparative study of family-based single common variant association tests, in which we simulated family data under different scenarios, compared and evaluated the Type I error, power and robustness of many statistics with different algorithms and implementations. Then, we discussed some applications of these compared statistics. In Chapter 3, we extended a gene-based kernel statistic for rare variants and binary traits to deal with family data, and evaluated its Type I error and power by simulation as well as compared it to other similar statistics. Then in Chapter 4, we discussed some advantages and disadvantages of existing methods and potential future work.

1.2 GENERAL BACKGROUND

1.2.1 Trait and marker

A trait is either a continuous or a binary expressed phenotype, which is controlled by a unobserved disease locus genotype; genetic markers are based on DNA polymorphisms. There are several different genetic markers. For example, there are single nucleotide polymorphisms (SNP), copy number variation (CNV) and restriction fragment length polymorphisms (RFLP). In Chapter 2, we have simulated one binary trait (e.g. affected/unaffected) and

one bi-allelic marker, which uses alleles to measure the polymorphism at a given locus on a pair of chromosomes; while in Chapter 3, we have simulated one binary trait and many genetic markers that within a selected genetic region.

1.2.2 Association analysis

In general, association analysis is, by applying an appropriate statistical test, trying to identify the relationship between a trait (an unobserved disease locus) and genetic markers. The purpose of association analysis is trying to provide genetic evidence of the etiology of a certain disease. Oftentimes, association analysis is also called gene mapping between a disease phenotype (trait) and marker genotypes.

1.2.3 Population and family data

There are generally two types of data that can be collected in order to identify the relationship between a trait and a marker using any statistic. The first one is population-based data, in which the individuals are randomly sampled from a huge population, thus assumed to be independent to each other; while the other one is family-based data, in which the individuals are family members, thus assumed to have correlations with each other within families. Using family-based data can guard against confounding factors such as population stratification, which means the population itself can be separated into different groups just by differences of allele frequencies between sub-populations due to different ancestry. In this case, allele frequency based association analysis would be confused by the confounding factor. In this dissertation, we focused on family-based data and discussed the applications in the presence of population stratification.

1.2.4 Linkage and association

There are basically two types of tests that can be applied for a single marker analysis. One is called linkage test, where linkage can be viewed as a measurement of the correlation between the pattern of marker inheritance and the pattern of trait inheritance (recombination events

during meiosis) in a long-range along a chromosome; The other one is called association test, where the association can be viewed as a relationship (linkage disequilibrium or LD) between a marker allele and a trait in a short-range along a chromosome.

Therefore, one can apply association test statistics on both population-based and family-based data, while one can only apply linkage test statistics on family-based data. In fact, in family-based data, as illustrated in Figure 1.1 for a bi-allelic marker with genotypes "A/A", "A/a" or "a/a", there would be both association and linkage if the targeted marker is related to the trait and has been passed from generation to generation; and there could be a linkage signal alone if the targeted marker is not related to the trait but has been passed from generation to generation. However, in family-based data, it is unlikely that there is an association signal alone because if a marker is very close to a disease-causing mutation, then they will always be inherited together.

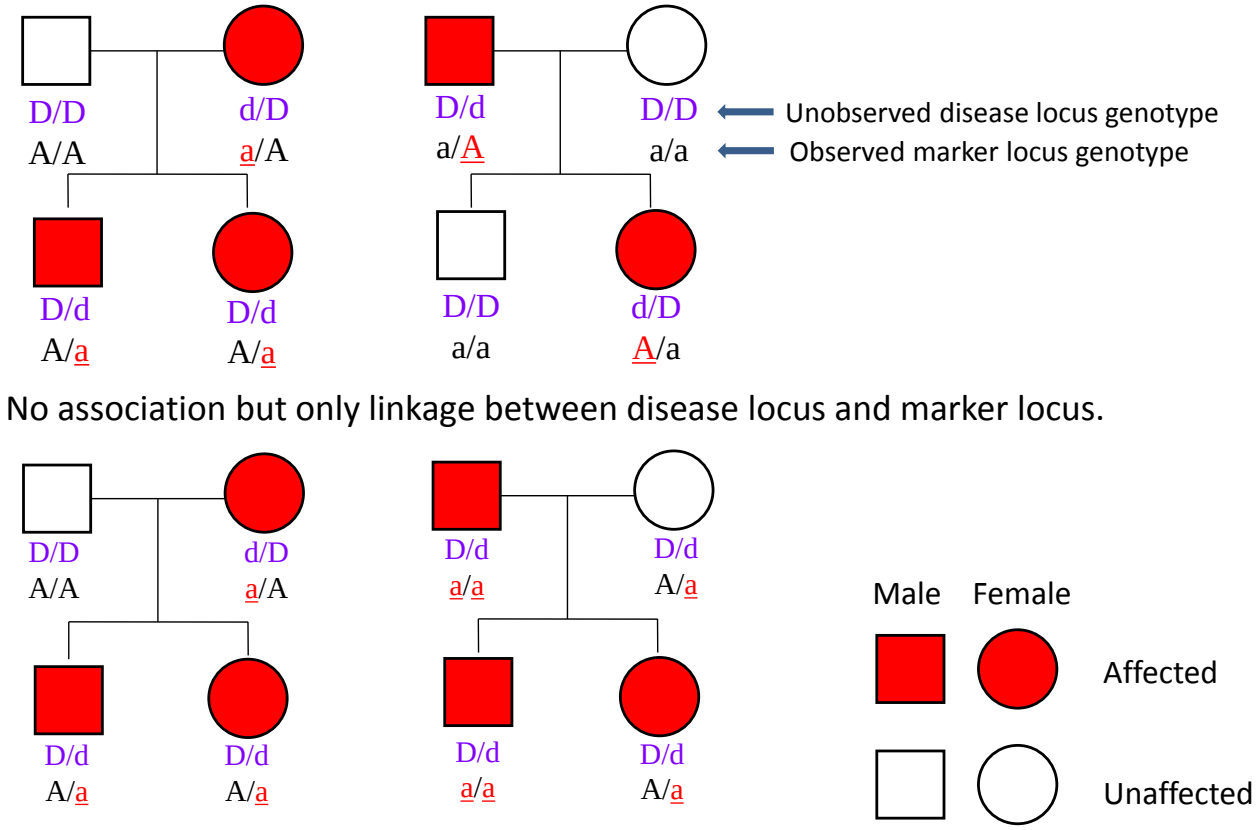


Figure 1.1: Example of linkage and association in family-based data.

1.2.5 Common and rare variants

For a bi-allelic marker, there are four different haplotypes: (0, 0), (0, 1), (1, 0) and (1, 1). Usually, we use '1' to represent the minor allele in haplotypes. These four haplotypes can be coded to genotypes by counting how many minor alleles are presented: 0: (0, 0), 1: (0, 1) or (1, 0) and 2: (1, 1). Minor allele frequency (MAF) can be calculated as the proportion of the minor allele in population. Based on MAF for each variant (marker) in the dataset, common variants usually are defined as the MAF greater than 5%, and we define those variants that have MAF smaller or equal to 5% as rare variants. Rare variants sometimes have larger effect size than common variants. Because of the small MAF, single variant analysis statistics will not be able to detect the association unless the sample size is large enough, so that people developed multi-variant (region-based) test statistics that identify the association between the traits and a genetic region, which contains a number of rare variants. Rare variants have not been as much researched as common variants. But as sequencing technology has improved, rare variants data are not as difficult to obtain as it was before. Therefore, it is necessary for statisticians to develop powerful statistics to do rare variants association analysis.

2.0 A SIMULATION-BASED COMPARATIVE STUDY OF FAMILY-BASED ASSOCIATION TESTS

2.1 MOTIVATION

The statistical genetics community has created a large number of different statistics for testing for association on family data. However, it is not necessarily clear which one of these would be best for a particular dataset. Furthermore, with the development of next generation sequencing technology, the pendulum is moving toward the increasing study of families, which will increase the need for family-based association analysis. It is essential to apply well-behaved, powerful, and robust statistics when analyzing family-based data. So it is important to evaluate, compare and summarize the statistical properties of commonly used family-based association test statistics when they are applied under different situations.

Some comparison studies of family-based association test statistics have been done: Chen et al. [2009] proposed a generalized disequilibrium test (GDT) and compared it to several other statistics such as the Family-based Association Test (FBAT) [Laird et al., 2000; Rabinowitz and Laird, 2000], and the Pedigree Disequilibrium Test (pdt) [Martin et al., 2001, 2000]. Their results showed that the GDT was the most powerful among those statistics. Hiekkalinna et al. [2011] compared their Pseudomarker statistics to some commonly used family-based association tests such as FBAT, MENDEL association test given linkage [Lange et al., 2001, 2005], QTDT [Abecasis et al., 2000a], TRANSMIT [Clayton, 1999] and UNPHASED [Dudbridge, 2008]. Pseudomarker was shown to have higher power than the other statistics.

In this Chapter, we compared a number of different association statistics (Table 2.1) on identical simulated datasets in a controlled environment to determine which ones are best

under which conditions. By simulating family data and varying the strength of association, family structures, and the disease model, we evaluated and compared the statistical properties such as Type I error, power and robustness of different association statistics. We explore, in a controlled comparison, which statistic is most powerful and robust on which kind of data, and how power changes as the simulation models change.

2.2 APPROACH

2.2.1 Null Hypothesis

When testing for association on family data, there are four different null hypotheses that could be tested. These null hypotheses include:

Null A: H_0: no association (D'= 0)

Null NL: H_0: no association and no linkage ($\theta = 0.5$ and D'= 0)

Null CL: H_0: no association given complete linkage (D'= 0 | $\theta = 0$)

Null AL: H_0: no association given no linkage (D'= 0 | $\theta = 0.5$)

where θ denotes the recombination fraction and D'measures the strength of association. Note that, although we focused on association analysis, we also included a few statistics that test Null L and Null LA. In this study, we evaluated all the statistics listed in Table 2.1 on the family-based data simulated under Null NL and Null CL listed above, as well as under appropriate alternative hypotheses. We did not simulate any data under Null AL, because it is unlikely that, two locus which are under strong linkage disequilibrium (high D') and very close on same chromosome, would segregate separately within a family. Table 2.1 also shows the null hypotheses of the association tests and defines short names for those statistics, which we will use when referring to them.

Table 2.1: Abbreviations, null Hypothesis and descriptions of all statistics evaluated in this study

Abbreviation	**Null Hypothesis**	**Description**
ALLELE_FREQ	No Association	Pedigree based allele frequency estimation [Boehnke, 1991] and chi-square test implemented in Mendel package [Lange et al., 2001].
AS|LINK	No Association Given Linkage	Test for association given linkage [Cantor et al., 2005] implemented in Mendel package [Lange et al., 2001].
CACO_FISHER CACO_ZMAX	No Association	Case Control test with Fisher and Z-max p-values as implemented in Mendel package [Lange et al., 2001].
FBAT	No Association And No Linkage	Family-based association test [Laird et al., 2000; Laird and Lange, 2006; Rabinowitz and Laird, 2000].

Table 2.1 Continued

Abbreviation	**Null Hypothesis**	**Description**
FBAT_e	No Association Given Linkage	FBAT with empirical variance estimator [Laird et al., 2000; Laird and Lange, 2006; Rabinowitz and Laird, 2000].
g_1tdt	No Association And No Linkage	TDT extension that allows one un-typed parent in each family [Sun et al., 1999] implemented in GDT package [Chen and Abecasis, 2007; Chen et al., 2009].
g_gee1	No Association	Generalized Estimating Equation with independent working correlation implemented in GDT package [Chen and Abecasis, 2007; Chen et al., 2009].
g_mqls	No Association And No Linkage	MQLS [Thornton and McPeek, 2007] implemented in GDT package [Chen and Abecasis, 2007; Chen et al., 2009].

Table 2.1 Continued

Abbreviation	**Null Hypothesis**	**Description**
g_pdt	No Association And No Linkage	Pedigree Disequilibrium Test [Martin et al., 2001, 2000] implemented in GDT package [Chen and Abecasis, 2007; Chen et al., 2009].
g_qlsw	No Association And No Linkage	Quasi-likelihood score test [Bourgain et al., 2003; McCullagh and Nelder, 1989] implemented in GDT package [Chen and Abecasis, 2007; Chen et al., 2009].
g_tdt	No Association And No Linkage	Transmission Disequilibrium Test implemented in GDT package [Chen and Abecasis, 2007; Chen et al., 2009].
GC1, GC2	No Association And No Linkage	Gamete Competition with preset (GC1) or estimated (GC2) allele frequencies [Sinsheimer et al., 2000, 2001] as implemented in Mendel [Lange et al., 2001].

Table 2.1 Continued

Abbreviation	**Null Hypothesis**	**Description**
GC1CT, GC2CT	No Association And No Linkage	GC1, GC2 with Complementary Transmission option.
GDT	No Association And No Linkage	Generalized Disequilibrium Test [Chen and Abecasis, 2007; Chen et al., 2009].
GEE_ind GEE_ex	No Association	Generalized Estimating Equation with independent (ind) or exchangeable (ex) working correlation implemented in R package "GEE".
IQLS	No Association And No Linkage	Incomplete-Data quasi-likelihood score test [Wang and McPeek, 2009].
LME	No Association	Generalized linear mixed model implemented in R package "MASS" [Venables and Ripley, 2002].
Mendel_TDT	No Association And No Linkage	Transmission Disequilibrium Test [Spielman et al., 1993; Terwilliger and Ott, 1992; Lazzeroni and Lange, 1998] as implemented in Mendel [Lange et al., 2001].

Table 2.1 Continued

Abbreviation	**Null Hypothesis**	**Description**
MM1	No Association	Polygenic model based score test implemented in R package GenABEL [Aulchenko et al., 2007].
MQLS_e	No Association And No Linkage	More Powerful Quasi-likelihood Score test [Thornton and McPeek, 2007] implemented by Liang (`www.sph.umich.edu/csg/liang/MQLS`).
MQLStest_r	No Association And No Linkage	More Powerful Quasi-likelihood Score test [Thornton and McPeek, 2007].
MQLStest_caco	No Association And No Linkage	Case-control corrected quasi-likelihood score test [Bourgain et al., 2003; Thornton and McPeek, 2007].
PENE	No Association	Likelihood ratio test based on Generalized Linear Penetrance Model [Lange et al., 2005] implemented in Mendel [Lange et al., 2001].

Table 2.1 Continued

Abbreviation	**Null Hypothesis**	**Description**
PMDom_L	No Linkage	
PMRec_L	No Linkage	
PMMbase_L	No Linkage	
PMDom_L|LD	No Linkage Given Association	
PMRec_L|LD	No Linkage Given Association	
PMMbase_L|LD	No Linkage Given Association	
PMDom_LD|L	No Association Given Linkage	
PMRec_LD|L	No Association Given Linkage	Psuedomarker(PM) [Göring and Terwilliger, 2000; Hiekkalinna et al., 2011, 2012] wiht penetrance model dominant (Dom) and recessive (Rec) or model based (Mbase).
PMMbase_LD|L	No Association Given Linkage	
PMDom_LD|NL	No Association Given No Linkage	
PMRec_LD|NL	No Association Given No Linkage	
PMMbase_LD|NL	No Association Given No Linkage	
PMDom_LDL	No Association And No Linkage	
PMRec_LDL	No Association And No Linkage	
PMMbase_LDL	No Association And No Linkage	

Table 2.1 Continued

Abbreviation	**Null Hypothesis**	**Description**
poGDT	No Association And No Linkage	Generalized Disequilibrium Test but only examines discordant parent-offspring pairs.
QTDT_ad	No Association	General version of TDT Test use all available genotypic information from every individual, implemented in QTDT package [Abecasis et al., 2000a,b; Fulker et al., 1999].
QTDT_am	No Linkage	Monks model [Monks et al., 1998] implemented in QTDT package [Abecasis et al., 2000a,b; Fulker et al., 1999].
QTDT_ar	No Association And No Linkage	Rabinowitz model [Rabinowitz, 1997] implemented in QTDT package [Abecasis et al., 2000a,b; Fulker et al., 1999].

Table 2.1 Continued

Abbreviation	**Null Hypothesis**	**Description**
Transmit	No Association	TRANSMIT tests for association between genetic marker and disease by examining the transmission of markers from parents to affected offspring [Clayton, 1999].
Transmit_r	No Association	TRANSMIT with robust variance estimator [Clayton, 1999].
WQLS_r	No Association And No Linkage	Quasi-likelihood score test [Thornton and McPeek, 2007; Bourgain et al., 2003; McCullagh and Nelder, 1989].

2.2.2 Simulation Description

In order to thoroughly compare the statistics, the data were simulated under several different simulation scenarios (Table 2.2). To mimic different family structures in real data, we simulated two different family structures: two-generation and three-generation. Note that in Figure 2.1, the family structures are examples, while in simulation, the number of offspring for each generation in each family were randomly generated according to a negative binomial distribution with dispersion parameter 3.84 and probability 0.93.

Table 2.2: Simulation Scenarios with Marker/Disease allele frequency = 0.2/0.2

Family Structure	**Number of families**	**Penetrance Model**	**Null NL, Null CL, Alternatives**
2 generation families (2gen)	80	Dom (0.05, 0.35, 0.35) Rec (0.05, 0.05, 0.35) Add (0.05, 0.175, 0.35)	Null NL: no linkage (θ = 0.5), no association (D'= 0), 1000 replications. Null CL: complete linkage (θ = 0), no association (D'= 0), 1000 replications. Alternatives: complete linkage (θ = 0), D'= (0.4, 0.5, 0.6, 0.7), 200 replications.
3 generation families (3gen)	25	Dom (0.05, 0.35, 0.35) Rec (0.05, 0.05, 0.35) Add (0.05, 0.175, 0.35)	
2 generation families with one untyped parent (2genUP)	80	Dom (0.05, 0.35, 0.35) Rec (0.05, 0.05, 0.35) Add (0.05, 0.175, 0.35)	
3 generation families with two grandparents untyped (3genUG)	25	Dom (0.05, 0.35, 0.35) Rec (0.05, 0.05, 0.35) Add (0.05, 0.175, 0.35)	
3 generation families with two grandparents and some parents untyped (3genUGP)	25	Dom (0.05, 0.35, 0.35) Rec (0.05, 0.05, 0.35) Add (0.05, 0.175, 0.35)	

For one dataset, we first simulated 80 two-generation families with both parents genotyped and phenotyped (fully-typed), which we named it as "2gen" for short, as shown in Figure 2.1. And then, we assigned fixed binary disease status (phenotypes) to one child in each two-generation family, which is equivalent to ascertain the families that contain one affected child from the population, and conditionally simulated traits for the remaining family members and genotypes for everyone at a two-allele marker with allele frequencies (0.8,

0.2) using the simulation program FastSLINK [Cottingham et al., 1993; Ott, 1989; Schaffer et al., 2011]. The disease allele frequencies are set as $p_{affected} = 0.2$, $p_{unaffected} = 0.8$. We used Mega2 [Mukhopadhyay et al., 2005] to transform the simulated data into nine different formats as required by the various analysis programs. For those statistics that require pre-specified marker allele frequencies, we used the allele frequencies estimated from Mendel package analysis option 6: "Allele Frequencies", which provides estimates of allele frequencies by using the pedigree information [Lange et al., 2001]; for those statistics that require a pre-set prevalence, we set it at 0.05; and for those statistics that require a pre-set penetrance model, we set it as: (0.05, 0.45, 0.90).

We also simulated scenarios with different family structures, and gave them short names for easy reference in the following parts of this chapter (Figure 2.1). In specific, we simulated scenarios where the dataset contains 80 two-generation families with one untyped parent (2genUP), which means that parent was neither genotyped nor phenotyped, and scenarios where the dataset contains 25 fully-typed three-generation families (3gen), then scenarios where the dataset contains 25 three-generation families with both grandparents untyped (3genUG), and finally, another set of scenarios where the dataset contains 25 three-generation families with grandparents and some parents untyped (3genUGP). Note that we assigned fixed binary disease status (phenotypes) to one child in the first two parent-children families in each three-generation family, which is equivalent to ascertain the families that contain two affected subjects, who belong to two separated families at the bottom generation, and then conditionally simulated disease status for the remaining family members. The sample size in each dataset was controlled at around 500.

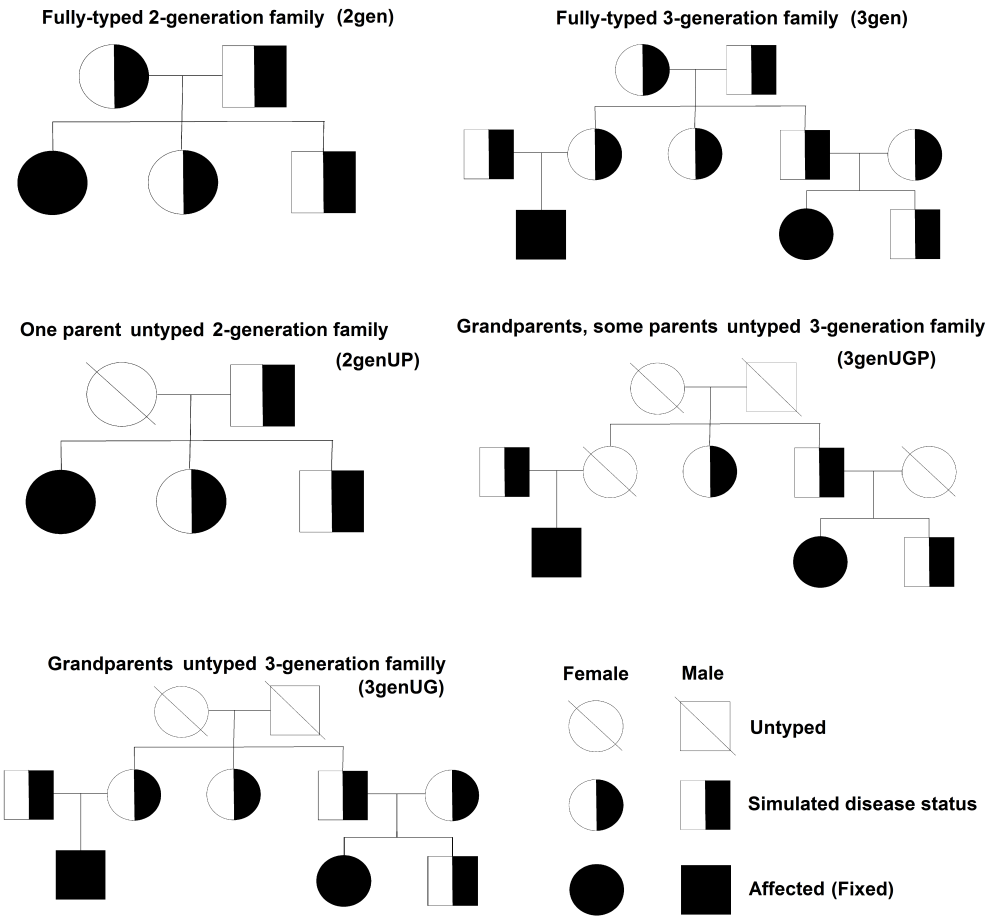


Figure 2.1: Examples of simulated family structures. The numbers of offspring are randomly generated from a negative binomial distribution.

To measure Type I error, we simulated 1,000 replicates. Within each replicate, first we simulated two datasets under two different null hypotheses: Null NL and Null CL, respectively. Second, we calculated p-values for each statistic under these two nulls. Third, after finishing the 1,000 replicates, for each statistic, we calculated the Type I error as the portion of the 1,000 P-values that were smaller than the pre-set threshold ($\alpha = 0.05$). For comparison purposes, based on 1,000 replicates, we calculated the boundaries of a 95% confidence interval of $\alpha = 0.05$: [0.037, 0.064] = $(0.05 \pm 1.96 \times \sqrt{(0.05 \times (1 - 0.05)/1000)})$ and defined three Type I error rates categories in terms of the boundaries of the 95% confidence interval. As illustrated in Table 2.3, if the estimated Type I error fell within the confidence interval, we labeled it as "Well-behaved". Otherwise, if the estimated Type I error fell in [0, 0.037) or (0.064, 1], we labeled it as "Conservative" or "High FP" (high false positive rate), respectively. And we plot the Type I error values in the columns "plot" in Table 2.3. Within each of those plots, there are five segments, each segment connects three points represents the three Type I error values under dominant, recessive and additive penetrance models, respectively. For these five segments, they represent five different family structures: 2gen, 3gen, 2genUP, 3genUG, 3genUGP, respectively starting from the left. Table A1 contains the Type I error values for all the statistics in Table 2.1 across all scenarios in Table 2.2. To measure power, 200 replicates were simulated for each of the alternatives in Table 2.2. The power is estimated as the fraction of p-values that are ≤ 0.05, based on 200 replicates simulated under each of the different simulation scenarios (Table 2.2). To measure robustness, we measured the behavior of our statistics in the presence of untyped individuals (Figure 2.1, and under additive, dominant and recessive penetrance models, where penetrance is the probability of being affected given a certain trait genotype (Table 2.2). A desirable statistic should have a consistently high power with a well-behaved Type I error across our simulated scenarios. In this study, after the simulation, we also clustered the statistics using Manhattan distances (absolute distance between two vectors) based on their p-values under Null NL across the scenarios where the data do not contain untyped individuals (we used R function 'hclust').

2.3 RESULTS

2.3.1 Clusters

The association test statistics we have compared in this study can be broadly classified into three categories: Transmission-based, Regression-based and Likelihood-based according to their basic characteristics. Transmission-based methods (e.g. Mendel_TDT, FBAT) usually construct the test statistics based on the count of alleles that are transmitted from parents to affected offspring. In other words, these statistics condition on parental genotype and thus robust to confounding factors such as population stratification. Regression-based methods (e.g. LME) usually construct regression models between trait and marker while adjusting for the correlation structure induced by the family structure. These methods themselves cannot effectively control for population stratification, but they can adjust for potential covariates, so that one can put in a principle component as a covariate that adjusts for population stratification. Likelihood-based methods can be separated into likelihood ratio tests (e.g. Pseudomarker tests), which construct the likelihood based on disease phenotype and marker genotype, then test the null hypotheses using likelihood ratio tests; and quasi-likelihood score tests (QLS, e.g., MQLStest), which build the likelihood based on allele frequencies and test the null hypotheses using a score test. Most of these methods are powerful but cannot control for population stratification except the one that has been recently developed: Roadtrips [Thornton and McPeek, 2010], which can control for population stratification.

When we cluster our family-based association statistics based on their Type I errors under Null NL, they fall into groups that reflect their underlying assumptions, algorithms and characteristics. Figure 2.2 shows that Group A contains regression-based statistics and MQLStest_caco, Group B contains Pseudomarker statistics, which are all based on likelihood ratio tests. Group C contains quasilikelihood-based case-control statistics and Group D contains transmission-based statistics.

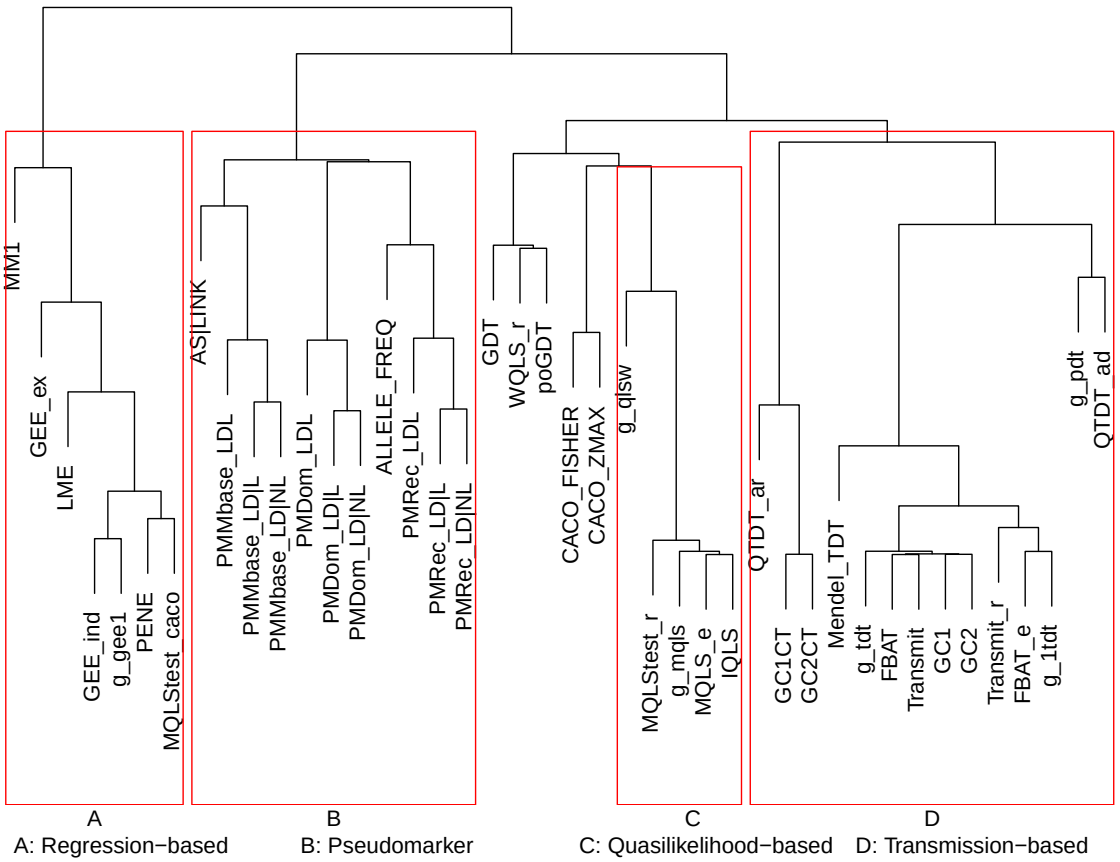


Figure 2.2: Hierarchical clustering plot based on Manhattan distance of p-values under Null NL across fully typed family structures and penetrance models.

2.3.2 Type I error

The statistics in our study also can be categorized into four groups according to their null hypotheses as we introduced in section 2.2.1 and in Table 2.1. Within each group, by comparing their Type I error behaviors, power and robustness, we would like to select the statistic which has well-behaved Type I error, high power and good robustness. In the following sections, we first compared Type I error behaviors of the statistics within each of the null hypotheses, and then we dropped the statistics that had inflated Type I error (H) or more than one deflated (C) Type I errors and compared power only for the ones that had well-behaved Type I errors.

Table 2.3 shows that, in the group of statistics that test for association, ALLELE_FREQ, which is the Mendel association test based on allele frequencies estimated by maximizing the likelihood that takes pedigree structure into consideration, has well-behaved Type I error under Null CL, which means it can control for linkage, but it has one conservative and one inflated Type I error behavior under Null NL when there are untyped individuals.

Table 2.3: Counts of Type I error behavior and power across 15 different scenarios.

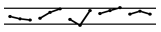
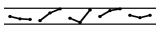
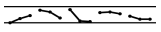
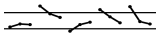
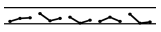
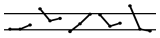
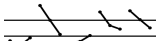
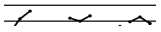
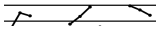
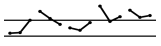
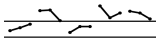
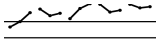
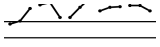
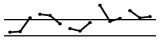
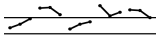
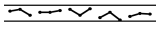
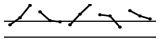
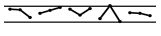
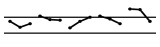
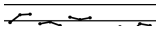
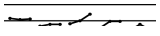
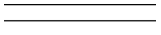
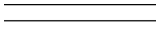
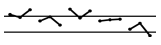
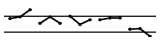
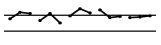
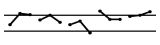
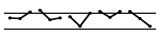
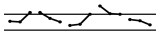
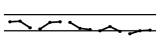
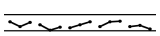
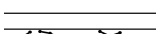
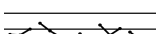
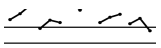
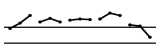
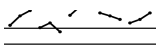
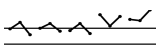
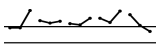
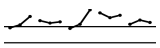
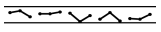
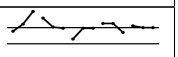
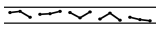
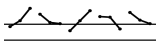
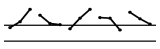
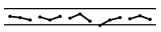
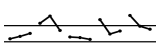
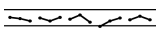
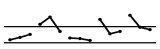
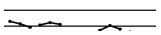
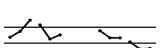
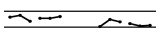
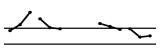
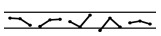
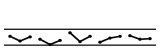
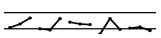
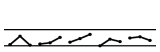
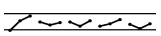
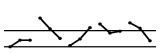
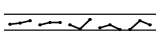
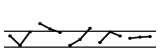
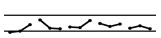
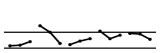
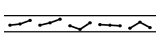
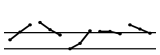
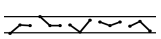
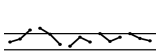
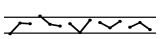
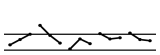
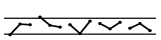
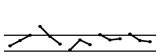
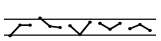
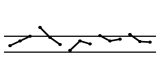
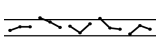
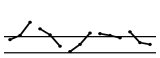
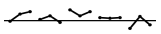
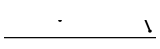
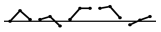
	No Linkage				Complete Linkage			
Statistics	C	W	H	plot	C	W	H	plot
Test for association (Null A)								
ALLELE_FREQ	1	13	1		0	15	0	
CACO_FISHER	0	15	0		1	11	3	
CACO_ZMAX	0	15	0		2	11	2	
PENE	7	8	0		5	6	4	
LME	6	6	3		5	7	3	
GEE_ind	0	8	7		0	6	9	
GEE_ex	0	2	13		0	1	14	
g_gee1	0	7	8		0	7	8	
Transmit	0	15	0		0	4	11	
Transmit_r	0	14	1		0	11	4	
QTDT_ad	10	5	0		10	5	0	
MM1	0	0	15		0	0	15	
Test for association in the absence of linkage (Null AL)								
PMDom_LD|NL	1	10	4		1	13	1	
PMRec_LD|NL	0	9	6		1	10	4	
PMMbase_LD|NL	0	10	5		0	11	4	
Test for association in the presence of linkage (Null CL)								
FBAT_e	2	13	0		0	15	0	
AS|LINK	15	0	0		11	4	0	
PMDom_LD|L	0	2	13		0	2	13	
PMRec_LD|L	0	1	14		0	6	9	
PMMbase_LD|L	0	3	12		0	2	13	
Test for association or linkage (Null NL)								
QTDT_ar*	2	10	0		0	9	3	

Table 2.3 Continued

Statistics	No Linkage				Complete Linkage			
	C	W	H	plot	C	W	H	plot
FBAT	0	15	0		0	8	7	
GC1	0	15	0		0	3	12	
GC2	0	15	0		0	3	12	
GC1CT	1	14	0		0	11	4	
GC2CT	1	14	0		0	11	4	
Mendel_TDT*	6	6	0		5	8	2	
g_tdt*	0	12	0		3	6	6	
g_1tdt	1	14	0		0	15	0	
g_pdt	4	11	0		0	15	0	
GDT	1	14	0		0	9	6	
poGDT	0	15	0		0	12	3	
MQLStest_caco	1	14	0		0	13	2	
WQLS_r	0	15	0		0	6	9	
MQLStest_r	1	14	0		0	13	2	
MQLS_e	1	13	1		0	12	3	
IQLS	1	13	1		0	12	3	
g_mqls	1	13	1		0	12	3	
g_qlsw	0	13	2		0	7	8	
PMDom_LDL	0	4	11		0	0	15	
PMRec_LDL	0	3	12		0	0	15	
PMMbase_LDL	0	9	6		0	6	9	

Note: C: Conservative, W: Well-behaved, H: High False Positive; Blue colored values are power, others are Type I errors. "plot": five segments correspond to family structures: 2gen, 3gen, 2genUP, 3genUG, 3genUGP, respectively. Each segment connects three points correspond to Type I error values under dominant, additive, recessive penetrance models, respectively. *: These statistics do not run in 2genUP families.

CACO_FISHER and CACO_ZMAX, which are case-control tests based on contingency table, are well-behaved under Null NL, but they have more than one inflated Type I errors under Null CL. PENE, which is Mendel penetrance based association test, has too many deflated Type I errors in two-generation families (2gen, 2genUP) and some inflated Type I errors in three-generation families (3gen, 3genUG, 3genUGP) under Null CL. LME (Generalized Linear Mixed Model) has too many inflated or deflated Type I errors. Note that if the segments are missing in those plots, that means they are outside the plotting area. GEE_ind (Generalized Estimating Equation with an independent working correlation implemented in R-package 'GEE') and g_gee1 (GEE with independent working correlation implemented in 'GDT' package) are essentially equivalent. We can see that all GEE statistics have inflated Type I errors in three-generation families and GEE_ex (Generalized Estimating Equation with exchangeable working correlation implemented in R-package 'GEE') has inflated Type I error in all five family structures. Transmit are well-behaved under Null NL, but has inflated Type I error in the presence of complete linkage. Transmit_r is Transmit with robust variance estimator, which enables the use of Transmit on families with more than one affected offspring, and even in the presence of linkage. Transmit_r shows better behaviors than Transmit under Null CL, but it still have inflated Type I error especially in three-generation families with family structure 3genUGP. QTDT_ad has heavily deflated Type I errors. While the polygenic model based score test (MM1) has heavily inflated Type I error behavior. Thus, in this group, we select ALLELE_FREQ and evaluate its power behavior.

The group of statistics that test for association in the absence of linkage are all Pseudo-marker statistics. We can see that these statistics have inflated Type I error. However, since strong association with a nearby causative disease locus usually implies linkage in family data, so we do not simulate any dataset with association but no linkage. Thus, we do not further evaluate the Type I error or power behaviors of PMMbase_LD|NL, PMRec_LD|NL or PMDom_LD|NL.

In the group of statistics that test for association in the presence of linkage, Table 2.3 shows that FBAT_e, which is FBAT (Family-Based Association Test) with empirical variance estimator has well-behaved Type I error under Null CL, but has conservative Type

I error under Null NL in 3genUGP families. Pseudomarker statistics have inflated Type I error. While, with the same setting of the penetrance model as the one for model-based Pseudomarker statistics, AS|LINK, which is the Mendel likelihood ratio test for association given linkage, has deflated Type I error. Thus, in this group, we select FBAT_e and evaluate its power behavior.

In the group of statistics that tests the null of no association and no linkage, under Null CL (D'= 0 and θ= 0), these statistics are expected to generate significant p-values to reject their null hypotheses, thus, the simulation here measures their power to detect linkage. Note that, although differs by implementation, MQLStest_r, MQLS_e g_mqls and IQLS are essentially equivalent. And QTDT_ar, Mendel_TDT and g_tdt do not run in 2genUP families. Under Null NL, Table 2.3 shows that, most of these statistics have well-behaved Type I errors, except for some Quasi-likelihood tests namely MQLS_e, IQLS, g_mqls g_qlsw and Pseudomarker statistics that jointly test for association and linkage. But note that the one inflated Type I error for those quasilikelihood-based statistics are very close to the boundary. QTDT_ar has deflated Type I errors in 3genUGP families, Mendel_TDT has deflated Type I errors when the family contains untyped parents or grandparents. Recall that we only select the statistics that have zero inflated Type I error (H) or less than two deflated (C) Type I errors. Thus, in this group, we select FBAT, GC2, GC2CT, g_tdt, g_1tdt, GDT, poGDT, MQLStest_caco, WQLS_r and MQLStest_r. Note that GC1 and GC2 have very similar behaviors, so we just select GC2 and drop GC1. Also, we select GC2CT and drop GC1CT since they have very similar behaviors. The actual values of Type I error are contained in Table A1.

2.3.3 Power

Recall that we dropped the statistics that had inflated Type I error behavior and more than one deflated Type I error count in the absence of untyped individual (Table 2.3). Thus, we have the following statistics left: ALLELE_FREQ, FBAT_e, FBAT, GC2, GC2CT, g_tdt, g_1tdt, GDT, poGDT, MQLStest_caco, WQLS_r and MQLStest_r.

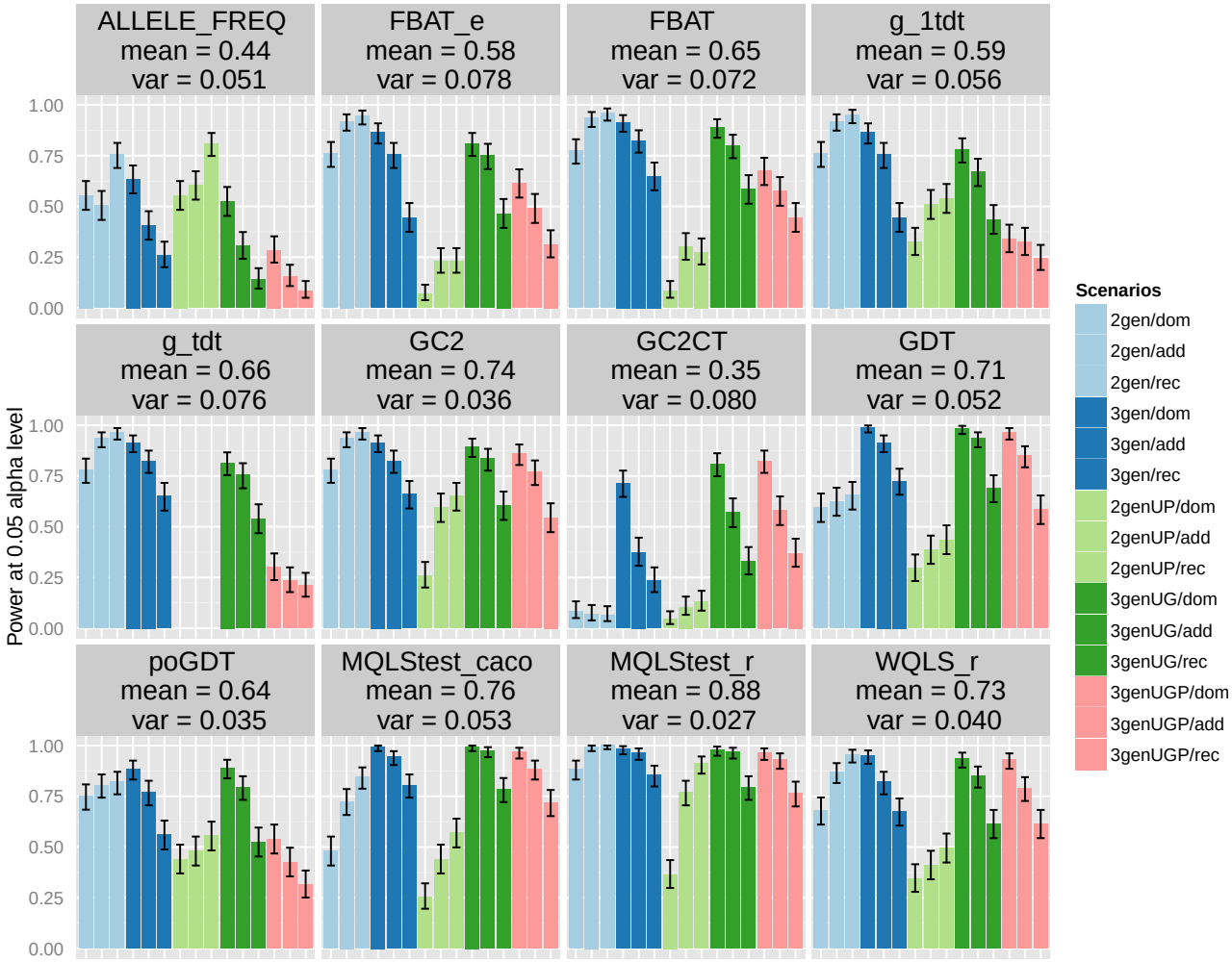


Figure 2.3: Power of selected statistics with well-behaved Type I error at 0.05 alpha level, with D'= 0.6, across all simulated scenarios. The bars within each statistic are ordered according to the legends. g_tdt does not work on 2genUP families. Error bars are 95% confidence intervals calculated based on 200 replicates.

Figure 2.3 shows the power of the selected statistics when D'= 0.6 across all the scenarios in Table 2.2. Note that ALLELE_FREQ, FBAT_e and the remaining statistics are testing three different null hypotheses. We can tell that the two statistics have similar power if their CIs are overlapping. We group the power bars in Figure 2.3 by statistics, which enables us to quickly tell which statistic has consistently high power across different scenarios. Figure 2.3 shows that ALLELE_FREQ, which tests for association, has medium power (around 50%) in 2gen and 2genUP families, but suffers power loss in 3gen, 3genUG and 3genUGP families, especially under recessive penetrance models. FBAT_e, which tests for association given linkage, has high power in 2gen families, but suffers power loss in 2genUP. It has medium power in 3gen, 3genUG and 3genUGP families, but has consistently low power under recessive penetrance models in those families. The other selected statistics test for association or linkage. Figure 2.3 shows that FBAT has similar power behaviors with FBAT_e although they have different null hypotheses. g_1tdt has high power in 2gen families and medium power in 3gen and 3genUG families, especially under recessive penetrance model. It has low power in 2genUP and 3genUGP families. g_tdt does not run in 2genUP families due to one untyped parent. But it has high power in 2gen and 3gen families, medium power in 3genUG families, and low power in 3genUGP families. GC2 has medium to high power in all scenarios, except for 2genUP under the dominant penetrance model, for which it has low power. GC2CT has low power in 2gen and 2genUP families, but has medium power in 3gen, 3genUG and 3genUGP families. GDT has high power in 3gen, 3genUG and 3genUGP families, but only has medium power in 2gen and 2genUP families. poGDT has similar power behaviors to GDT except in 3genUGP families, but with smaller variations across different scenarios. Among the quasilikelihood-based statistics, they have similar power behaviors, but MQLStest_r has, in average, the highest power. Note that, in all three-generation scenarios (3gen, 3genUG, 3genUGP), statistics have highest, medium, lowest power under dominant, additive, recessive penetrance models, respectively. However, in most two-generation scenarios (2gen, 2genUP), this order is reversed.

2.4 DISCUSSION

In this study, we evaluated several family-based association test statistics (referred by the abbreviations in Table 2.1) by applying them on simulated family data varying family structures (Figure 2.1), inheritance models, presence or absence of linkage and genotype/phenotype status of the parents (Table 2.2). We would like to find a statistic that has well-behaved Type I error and high power as well as robustness to untyped parents, different family structures and underlying penetrance models. There are four different null hypotheses (Section 2.2.1) that were tested by these family-based statistics. As it may be confusing and misleading to compare statistics that test different null hypotheses, we compare statistics that are testing the same null hypothesis.

Figure 2.2 shows the groups of statistics that are clustered together based on their p-values under Null NL across different scenarios. We have calculated Manhattan distance, which measures the absolute distance (L1 norm) between two vectors; we also have found similar results (Figure A1) by using Euclidean distance, which measures the squared distance (L2 norm) between two vectors. Note that these groups are not clustered based on power because any two statistics, although different in algorithm, could be clustered together if they have similar power behaviors, which influences the distance between two vectors of p-values. However, under Null NL, the dataset contains neither association nor linkage, so that the behavior of the p-values are purely depend on the underlying algorithm of the statistics. Someone may argue that, for each statistic, p-values under null shall follow a zero-one uniform distribution if the statistic is well calibrated, so that each replicate of simulation would generate a p-value randomly from a zero-one uniform distribution. This is true if we look at the distribution of p-values over a number of replicates in simulation. However, here the comparisons are among statistics. Two well calibrated statistics could generate two different p-values when they are applied to the same dataset that is simulated under null, where the difference between these two p-values implies the difference between the two statistics, which could have different implementations or different underlying algorithms. In other words, statistics have different Type I error behavior due to their algorithms. And Type I error is a function of p-values under null in simulation studies, thus, it is interesting and

reasonable to cluster different statistics based on their p-values and the results in Figure 2.2 and Figure A1 reflect the underlying algorithm of the statistics. In Table 2.3, we have summarized the Type I error behaviors for all statistics in Table 2.1, and we also listed the Type I error values in Table A1. In the group of statistics that test for no association, the inflated Type I error for CACO_FISHER and CACO_ZMAX implies that permutation of case-control labels within each family isn't sufficient to attain good Type I error rates when the family structure becomes complicated. But, the Type I error behaviors do not change when the data contain untyped individuals, which means these two methods are robust to untyped individuals. For LME, it is interesting that even with pre-specified correlation structure, which is the expected kinship matrix calculated based on family structure, LME still cannot control Type I error in three-generation families, which may imply that logit link cannot model the kinship correlation very well for a binary trait, single marker analysis. For GEE-based statistics, it looks like their Type I error behaviors do not get influenced by the untyped individuals since GEE-based methods allow missing data. However, we can tell that the exchangeable working correlation is inappropriate to model the family structure. Even with independent working correlation, Type I errors are inflated in three-generation families although GEE-based methods are robust to mis-specifications of correlation structure. Since one can calculate expected kinship correlations given the correct family structure, one may use a function of kinship coefficients as the working correlation in order to better control the Type I error behaviors of GEE-based statistics. Transmit and Transmit_r are robust to different family structures, but not robust to different penetrance models, and Transmit_r cannot control for complete linkage when there are untyped people.

In our simulation, Pseudomarker statistics that test for association are not well-behaved in Type I error. But in other simulation scenarios, one may find Pseudomarker statistics are well-behaved [Hiekkalinna et al., 2011]. However, even when they are well-behaved, the problem is that, by using "-all" option, the Pseudomarker program is able to calculate all 15 Pseudomarker statistics all at once; if we report the smallest p-value among these 15 statistics, then there is a potential multiple testing problem. So we should choose only one of these 15 statistics in advance, and use it through all the analysis, to avoid this problem. Also, Hiekkalinna et al. [2011] compared Pseudomarker statistics to FBAT, AS|LINK,

Transmit, and QTDT. They find that Pseudomarker statistics, Transmit, and QTDT are well-behaved, which are different from ours; AS|LINK is badly deflated, which is similar to what we observed in Table 2.3. Note that, in their study, not only AS|LINK and FBAT are set with recessive and dominant penetrance models, but also Pseudomarker recessive and dominant statistics are selected respectively to match the underlying true (simulated) penetrance models. However, in real data, usually researchers have very limited information about the underlying penetrance model. While, in our study, in order to fairly compare each statistic to others, we grouped the statistics by their null hypothesis, and compared them across all different scenarios to evaluate the robustness to different penetrance models while assuming the penetrance models are unknown.

In the group of statistics that test for no association or no linkage, most of them have well-behaved Type I error. Note that we include GDT and poGDT into the group of statistics that test for association or linkage, the reason is because our results show that the presence of linkage will inflate GDT statistic even when there is no association (Table 2.3). Although Chen et al. [2009] also recognized this inflation of the GDT test statistic due to the presence of linkage in their study, and suggested using local identity-by-descent (IBD) estimates instead of kinship-derived IBDs to correct for this, they did not clearly define the null hypothesis of GDT as no association and no linkage. Also in their study, MQLStest_r was not compared to GDT, among the statistics that were compared, GDT shows well-behaved Type I error at 0.01 alpha level, which are similar to the results in our study.

2.4.1 Population stratification

One important issue for association analysis is that when there are population substructures, for example, population stratification, marker allele frequency could be different in the sub-population level, which will introduce bias for some statistics whose algorithms depend on marker allele frequencies even though the study design is family-based. In our study, Quasilikelihood-based statistics, for instance, MQLStest_r, which tests the null hypothesis of no association and no linkage, is a desirable statistic: it has well-behaved Type I error and consistently high power. Moreover, by using local kinship coefficients, it also can be used to

test the null hypothesis of no association given linkage [Thornton and McPeek, 2007]. But it is not robust to population stratification because it compares the difference of allele frequencies between cases and controls while using kinship matrix that is either calculated from family structure or estimated from the genotype data (posterior) to control for relatedness among individuals. However, Thornton and McPeek [2010] has extended MQLStest_r to be also robust to population stratification, pedigree errors and unknown pedigree structures by constructing an estimator of kinship matrix from genome-screen data. This extended statistic is called ROADTRIPS, it assumes the correlation structure is the same across markers. In their study, ROADTRIPS has been compared to FBAT, MQLStest_r and MQLStest_caco on simulated data in the presence of population stratification. Their results have showed that, in the absence of population admixture, ROADTRIPS and MQLStest_r have similar power, but ROADTRIPS has better Type I error behavior and power in the presence of population admixture whereas FBAT has well-behaved Type I error but very low power. Besides Quasilikelihood-based statistics, Pseudomarker statistics are not robust to population stratification, either. Pseudomarker statistics construct a likelihood and maximize it over marker allele frequencies, which could be different among sub-populations.

Transmission-based statistics condition on observed parental genotypes, which eliminates nuisance parameters such as marker allele frequencies. Thus, these statistics should be able to control for population stratification. Although GDT, poGDT, GC1CT, GC2CT and QTDT_ar are not clustered together with Transmission-based statistics group in Figure 2.2, they are all Transmission-based statistics and robust to population stratification. Thus, we can also apply GC2 to test the null hypothesis of no linkage and no association in the presence of population stratification. However, the problem for GC2 is that when the data contain untyped individuals, it is no longer immune to population stratification because it fills in missing allele according to the population frequencies [Sinsheimer et al., 2000]. Also, similar to Mendel_TDT, when one applies them to trios, they test the null hypothesis of no linkage or no association instead of no linkage and no association [Lange et al., 2005]. Note that Mendel_TDT and g_tdt do not work on two-generation families with one untyped parent, in which case, g_1tdt, which can handle families with one untyped parent, would be a good alternative option. Among the rest of the statistics in Figure 2.3, GDT has high

power and well-behaved Type I error in Table 2.3, where poGDT is more robust to different scenarios, but less powerful than GDT. Note that GDT and poGDT are robust to population stratification and also can handle missing data well [Chen et al., 2009]. Although LME and GEE_ind, which test for association, show inflated Type I error in three-generation families, it can incorporate covariates in the model to control for population stratification, for example, by using PCA method [Price et al., 2006]. Chen et al. [2011] has compared LME, LMEBIN, which treats binary traits as continuous, and GEE with several different working correlation structures and variance estimators. Note that the LME and LMEBIN in their study have been carried out by using R package 'LME4', in which it is very difficult, if not impossible, to specify the correlation structure. In their study, the simulated dataset contains families with fixed three-generation family structure. When the prevalence is set at 0.05, GEE_ind and GEE_ex have showed deflated Type I error behavior, which is different from what we observed. This may be due to the fact that the family structure in their study is fixed, while in our study, every family has a different number of offspring. In the Mendel package, ALLELE_FREQ tests for association, but it should also be able to control for population stratification because it estimates marker allele frequency by using MLE of a likelihood that contains transmission probabilities, and then constructs a chi-square test of homogeneity to test for association. However, it only has medium power and robustness, which would not make it the first choice of statistics in real data analysis, but one can use it to estimate marker allele frequencies that can then be used by other statistics that require pre-set allele frequencies, which is what we have done in our simulation.

2.4.2 Mis-specified family structure

Besides the issue of population substructure, there are issues of mis-specified family structures, which could affect all the transmission-based statistics and all the statistics that use kinship coefficients in their algorithms. In this case, GEE_ind, CACO_FISHER or CACO_ZMAX could be the choice of statistics. GEE_ind uses independent working correlations, CACO_FISHER and CACO_ZMAX do permutations within each family, but their Type I errors in our simulation are slightly inflated (Table 2.3). Another potential problem

for GEE is convergence. However, we have encountered this problem only when we use exchangeable working correlations (GEE_ex), which was also an issue that has been found in Slager et al. [2003], but not if we use independent working correlations.

2.4.3 Test for association in the presence of linkage

In real data analysis, usually one can measure the strength of linkage by calculating the LOD score of a genetic region, which contains a number of genetic markers. Then the researchers would have prior knowledge of the strength of linkage, and would like to test for association within the linkage region. In this case, we can apply FBAT_e, which has well-behaved Type I error under all the scenarios, but has medium power and low robustness to different scenarios.

If the data contain trios without untyped individual, but have population stratification problem, one can also apply GC2 or Mendel_TDT in the region where there is linkage, as they test the null hypothesis of no linkage or no association. Thus, it requires both linkage and association to reject the null hypothesis. However, if the data contain nuclear families with more than one affected offspring, GC2 and Mendel_TDT confound association with linkage. In our simulation, one can see that GC2 has 7% power when there is no association but complete linkage (Table A1), compared to 80% power when there is both association (D'= 0.6) and complete linkage (Figure 2.3). In this case, another choice could be g_1tdt or g_pdt, which can handle families that have only one fully-typed parent. But they are less powerful than Mendel_TDT or g_tdt. Table 2.3 shows that, although g_1tdt and g_pdt test for association and linkage, under complete linkage, their power do not increase compared to their Type I error under no linkage. Power-wise, this is not a good thing because these two statistics are supposed to detect linkage, but, in other words, these two statistics are robust to linkage signal. Figure 2.3 shows their power when there is association (D'= 0.6) and complete linkage in the data, because they are not sensitive to linkage, their power are purely from the association signal. Thus, I believe, ambiguously, we can apply these two statistics on the region in the presence of linkage, but this needs to be verified by a thoroughly carried simulation study in the future.

2.4.4 Ascertainment criteria and study design

Also, in this study, we have simulated and selected the families with some ascertainment criteria. Specifically, we have ascertained those two-generation families that contain at least one affected child, and the three-generation families (Grandparents, parents, children) that contain one affected child in at least two parent-children families. The ascertainment procedure could introduce bias to population level parameter estimates [Clark et al., 2005; Siegmund and Langholz, 2002], for example, allele frequency, if not handle correctly in family-based data. In this study, we do not quite address this issue in the simulation, but according to the literature, "Deviance" option in PENE in Mendel package can be used to control for ascertainment procedure. And MQLStest_e and similar statistics assume the families are ascertained such that the data contain certain numbers of affected and unaffected individuals, this is probably one of the reasons that MQLStest_e is more powerful than others. Transmission-based statistics, for example Mendel_TDT, should be robust to the bias caused by ascertainment procedure since they eliminate the population level parameters by condition on parental genotypes. One possible study design to guard against ascertainment bias is to ascertain the sample according to one trait, and then analyze another trait that is not highly correlated with the ascertained one [Schifano et al., 2012]. But more simulations need to be carried out to evaluate the performance of this study design. Also, in real data analysis, one has to consider the family members may live at different regions in the world, such that their environment exposure factors can be quite different, Siegmund and Langholz [2002] has proposed a method to correct for this, but in our simulation study, we do not address this issue.

In Genome-wide association analysis, single marker test statistics are applied over a genetic region that contains several markers. Depends on the length of the genetic region, by calculating LOD score, one can see that the strength of linkage could be different over the region prior to the application of statistics. It is possible to divide the region into separate parts, among which the strength of linkage are different, then, one can apply the test statistics with different null hypotheses simultaneously to those separated genetic regions accordingly in order to obtain the best outcome. While then the problem becomes at where, what LOD

score value, we should separate the genetic region according to the strength of linkage, and what if some of the markers in separated regions are correlated. It would be interesting to simulate and compare this strategy versus the regular strategy in the future.

3.0 FAMILY-BASED RARE VARIANTS ASSOCIATION ANALYSIS FOR BINARY TRAITS

3.1 BACKGROUND AND MOTIVATION

Many population-based Genome-Wide Associations Studies (GWAS) have been carried out to look for common genetic variants that are associated with diseases. GWAS have successively identified more than 1,000 genetic loci, which are associated with many human diseases. In most GWAS, researchers usually focus on common genetic variants, which usually are defined as minor allele frequency (MAF) greater or equal to 5%, due to lack of power to detect rare variants (MAF $< 5\%$). Thus the diseases are assumed to be only associated with common genetic variants (MAF $\geq 5\%$). This forms an arbitrary assumption, which is called "Common Disease, Common Variant (CDCV)". However, there is another assumption called "Common Disease, Rare Variant (CDRV)", which argues that multiple rare genetic variants are not only associated with the diseases, but are also the major contributors, especially for complex and/or serious diseases. In fact, the common variants that are identified in GWAS often only explain at most 5% - 10% of the heritable component of a disease. The debates between these two assumptions still continue. Both of the assumptions have supportive evidence and have been discussed in Smith and Lusis [2002], Iyengar and Elston [2007] and Schork et al. [2009].

In population-based GWAS association studies, the data are collected from unrelated individuals, and in the past few years, family-based association analysis were not so popular due to higher costs of collecting related individuals compared to GWAS, in which it is cheaper and faster to collect unrelated individuals. However, as we come to the next-generation-sequencing era, faster and less expensive sequencing techniques have been developed. Thus,

collecting family-based genetic data has become cheaper and faster. Moreover, using family-based data can effectively control for the Type I error due to confounding factors (e.g. population substructure), which becomes more of an issue for rare variant association analysis because that rare variants can cause stronger stratification than common variants [Babron et al., 2012; Mathieson and McVean, 2012; Cheng and Chen, 2013; Mao et al., 2013; Jiang et al., 2013; He et al., 2014]. Although using the Principle Component Analysis (PCA) method can also control for Type I error due to population substructure, performing PCA is insufficient for admixed populations, even in the case of common variants [Liu et al., 2013a]. As we are entering the era of "Big Data", development of new study designs and new statistical tools for detecting rare genetic variants that are associated with complex diseases using family-based data are of significant importance.

3.1.1 Existing methods

Because of the low frequency (MAF $\leq 5\%$) of the rare genetic variants, those single marker statistical tests that we have compared in Chapter 2 are no longer powerful enough unless the sample size is very large and simply increasing sample size is difficult due to limited resources. To overcome this issue, researchers have developed some methods to aggregate information over a genetic region (e.g., a gene) in order to decrease the degrees of freedom of the test statistic. In general, there are three methods for aggregating information over a genetic region: one is the so-called 'burden' method, which, for each individual, sums the variant over all the markers in the genetic region to form an aggregated (increased) genetic variation signal; while the second one is the bi-directional (kernel) method that treats the coefficients of genetic markers as random effects within a mixed-model framework to decrease the degree of freedom of the test statistic. The third one is the Principle Component Analysis (PCA) method, which compares between the component of cases and the component of controls to decrease the degree of freedom of the test statistic. All of these three aggregating methods are implemented in many statistical tests, which will be explained further below.

3.1.1.1 Burden tests Although there are a lot of different implementations of burden tests, the concepts are quite similar: for each individual, aggregate the genetic variants within the genetic region of interest into one collapsed score, so that for all the individuals, there is a single column vector of collapsed scores. Currently, there are two different methods of collapsing, one is by counting, and the other is by dichotomizing, which means the collapsed scores are indicators (0 or 1) of whether the corresponding individual carries the rare allele(s) or not (Table 3.1). The collapsed scores are used for association tests, which would have higher power due to the stronger rare allele variation signal from collapsing over the region. The limitations of burden tests are that they have high power only when the most of the rare genetic variants within the region influence the traits in same direction [Schaid et al., 2013].

Table 3.1: Example of Burden test collapsing genotype

Person	Marker genotype	Counting	Dichotomizing
1	0 0 1 0 2 0 1 0 0	4	1
2	1 0 0 1 0 0 1 2 0	5	1
3	0 0 0 0 0 0 0 0 0	0	0
4	0 0 2 0 0 1 0 1 0	4	1

3.1.1.2 Bi-directional (Kernel) tests As mentioned above, burden tests only have high power under certain situations, which might not be satisfied when a genetic region contains a lot of non-risk rare genetic variants and/or the direction of effect of the variants could be more than one (risk and protective variants). In this case, there are some other statistics that have high power [Wu et al., 2011; Neale et al., 2011; Schaid et al., 2013]. These statistics are different from each other, but they all draw inference by using a variance component method. Specifically, for case-control data, one can compare the expected variance with the actual variance of the distribution of allele frequencies [Neale et al., 2011]. Or one can construct a random effect model by assuming that the marker coefficients (random coefficients)

are following a multivariate normal distribution with mean zero and a variance-covariance matrix ($\boldsymbol{\Sigma}_\beta$) that contains a variance component, then one can test whether the variance component is equal to zero or not, which is equivalent to test the null hypothesis of no association. For example, if one assumes independence among markers, the variance of each marker (every diagonal element in $\boldsymbol{\Sigma}_\beta$) can be expressed in the form of a multiplication of a weight (e.g. w_k for marker k) and a variance component (τ). Then, testing for association is equivalent to testing the null hypothesis that $\tau = 0$, such that all the marker coefficients are zero. For example, this test is illustrated below:

$$\overbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}^{Traits\ \boldsymbol{Y}_{n\times 1}} \quad \overbrace{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1t} \\ x_{21} & x_{22} & \cdots & x_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nt} \end{pmatrix}}^{Covariates\ X_{n\times t}} \quad \overbrace{\begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1m} \\ g_{21} & g_{22} & \cdots & g_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nm} \end{pmatrix}}^{Genotypes\ G_{n\times m}}$$

So, the model that we fit is

$$g(E[\boldsymbol{Y} \mid \boldsymbol{\beta}]) = \boldsymbol{\eta}^{\boldsymbol{\beta}}_{n\times 1} = \boldsymbol{\alpha}_0 + \boldsymbol{X}_{n\times t}\boldsymbol{\alpha}_{t\times 1} + \boldsymbol{G}_{n\times m}\boldsymbol{\beta}_{m\times 1} \tag{3.1}$$

where 'g()' is the link function, $\boldsymbol{\eta}^{\boldsymbol{\beta}}$ is the vector of linear predictors,$\boldsymbol{\alpha}_0$ is the intercept, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_t)^T$ is the coefficient for the fixed effect. And the random slope $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_m)^T \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$ where $\boldsymbol{\Sigma}_\beta = \tau \times \boldsymbol{W}$ and

$$\boldsymbol{W}_{m\times m} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_m \end{pmatrix}$$

where w_k represents the weight for β_k, k = 1, ..., m, and τ is a variance component. By doing this, the degrees of freedom is decreased (only one parameter τ is tested), so the power of the test would be increased. This algorithm has been implemented in "Sequencing Kernel Association Test (SKAT)" by Wu et al. [2011], in which it assumes $\sqrt{w_k}$ follows Beta(MAF_k; 1, 25) to increase the weights for rare variants, where MAF_k is the minor allele frequency for marker k.

3.1.1.3 Combined and PCA tests

In whole-genome sequencing studies, usually there are mixtures of risk and protective variants. In order to reach the optimal statistical power, one can use a linear combination of burden and kernel tests by giving them adaptively selected weights [Lee et al., 2012]. Besides combined tests, one can also adjust the p-values from applying multiple single-marker tests [Cheung et al., 2012; Fang et al., 2013; Lin et al., 2014], or construct the test by using a two-stage design [Zhu et al., 2010], or blocking approach [Turkmen and Lin, 2014] as well, which separates a genomic region into "independent" blocks and then aggregates the information over each block. Moreover, as proposed by Luo et al. [2011], one can use the PCA method to do dimension deduction and then conduct a rare variant association analysis, in which, they applied functional data analysis techniques to jointly test the association by testing the equality of two random functional principle components between cases and controls.

3.1.2 Familial correlation

In previous sections, statistical tests for rare genetic variants have been introduced. Among those tests, only a few that can analyze family-based data. In order to utilize the familial correlation contained in family-based data to obtain 1) a better control for Type I error due to population stratification, 2) a possibly increased statistical power from potential information provided by the familial correlation, and 3) a more sensible interpretation of the genetic association with the disease than using population (unrelated) data, people have developed some methods by either extending current family-based tests to test for rare variants, or adding familial correlation into current population-based rare variant tests.

One can extend those family-based common variants association tests to test for rare variants association. There is a popular family-based single marker association test, FBAT [Laird et al., 2000], which has been extended to a multi-marker gene-based version (FBAT-MM [De et al., 2013]). Specifically, the FBAT-MM test is a multivariate extension of the univariate FBAT test designed to simultaneously test a set of markers in a defined genetic region. Similar to burden tests, FBAT-MM assumes effects of the rare genetic variants are all in the same direction.

On the other hand, one can also add familial correlation to population-based rare genetic association tests. For burden tests, one can construct a linear mixed effects model by treating covariates and collapsed scores as fixed effects, and familial correlations as random effects [Chen et al., 2013]. For bi-directional tests, one can treat the genotypes (random coefficient) and familial correlation (random intercept) as random effects. Usually, one assumes the distribution of the familial correlation (random intercept) follows a multivariate normal distribution with mean zero, and a variance-covariance matrix that is proportional to the kinship coefficients of all the subjects [Chen et al., 2013; Oualkacha et al., 2013]. Although Ionita-Laza et al. [2013] have extend SKAT by conditioning the null distribution on parental genotypes, they did not use the whole information from considering familial correlations among other family relationships. For combined tests, one can also extend them by adding in familial correlations. For example, Jiang and McPeek [2014] have developed "Minimum P-value Optimized Nuisance parameter Score Test Extended to Relatives (MONSTER)" by adding familial correlations to the combined test "Optimal Unified test (SKAT-O)" [Lee et al., 2012]. For PCA based tests, Zhu and Xiong [2012] have proposed a method to extend population-based PCA tests to process family-based data by dividing the test statistic over a correction factor, which depends on kinship coefficients and the number of cases and controls.

3.1.3 Continuous traits versus binary traits

Currently, most of the powerful family-based rare genetic variants association statistics are only able to model the association between continuous traits and the targeted genetic region (e.g., "MONSTER" [Jiang and McPeek, 2014]). When the traits are binary, explicitly constructing the marginal likelihood function is difficult due to the evaluation of the multiple integrals over all sample subjects and random effects. Some approximation methods have been applied to solve this issue, for example, Laplace's method and quasi-likelihood are applied by Lin [1997] in variance component test, which is implemented first by Wu et al. [2011] in SKAT for both continuous and binary traits, but only for unrelated individuals, and then by Oualkacha et al. [2013] in "Adjusted Sequencing Kernel Association Test (ASKAT)" for related individuals (family-based data), but only for continuous traits.

By treating traits as fixed and genotypes as random, Schaid et al. [2013] has developed a statistic test that can model the association between binary traits and the genetic region while avoiding the problem of approximating multiple integrals for the marginal likelihood. But this method tends to be less efficient compared to those methods that treat traits as random variables, especially when the traits are continuous. Also, the method developed by Wang et al. [2013] is a GEE-based SNP set association test for continuous and discrete traits in family-based data. This method uses kinship coefficients as the correlation structure. An advantage is that it allows for the within-family correlation to be mis-specified. But in the paper [Wang et al., 2013], this method has not been applied to family-based data with a binary trait; in this study, we applied it to a binary trait and compared it to other statistics. In this chapter, we extend the statistic in Lin [1997] by deriving a generalized linear mixed model that contains familial correlation and a variance component score test for a binary trait, and building a new statistic to test for rare variants association on family-based data.

3.2 APPROACH

To carry out the family-based rare genetic variant association analysis for binary traits, we base our statistic on a generalized mixed effect model framework. In specific, we assume random coefficients for genetic effects, and random intercepts for the familial correlations. If the traits are continuous, this model would be exactly the same as the 'Adjusted Sequencing Kernel Association Test' (ASKAT) [Oualkacha et al., 2013]; but here we are focusing on binary traits, so we derive our statistic for binary traits based on the work of Lin [1997], who introduce a variance component score test that can be applied to unrelated individuals with binary traits and was proved to be locally most powerful [Wu et al., 2011]. We extend it by integrating the familial correlation into the statistic. Then, we draw inference by using quasi-likelihood and variance component score test under the null hypothesis of no association while controlling for familial correlations and covariates. The details are presented in the following subsections.

3.2.1 Proposed Method: Model setting

The proposed method is based on the kernel (bi-directional) test. Thus, similar to the settings in Model (3.1), let Y_i, i = (1, ... , n) represents a binary trait (0 or 1) measured on individual i, n is the total number of individuals. Let X_{ij}, j = (1, ... , t) represents the j^{th} covariate measured on individual i, t is the total number of covariates. Let G_{ik}, k = (1, ... , m) represents the genotypes, which are the count (0, 1 or 2) of the minor allele, on the k^{th} bi-allele marker measured on individual i, m is the total number of markers.

We assume that:

1) intercept: α_0

2) fixed effect coefficients: $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_t)^T$

3) random effect coefficients: $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_m)^T \overset{independent}{\sim} MVN(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$,

$$\text{where } \boldsymbol{\Sigma}_\beta = \tau \times \boldsymbol{W}_{m\times m} = \tau \times \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_m \end{pmatrix}$$

and w_k represents the weight (default weight is 1) for β_k, k = 1, ..., m, and τ is a variance component.

4) random intercepts P_i: according to Oualkacha et al. [2013], ignoring dominant effects, each subject has a random intercept for the familial correlation, and given a kinship matrix $\boldsymbol{\Phi}_{n\times n}$, the familial correlations for all the subjects are following a multivariate normal distribution: $\boldsymbol{P} = (P_1, P_2, ..., P_n)^T \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_P)$

$$\text{where } \boldsymbol{\Sigma}_P = \sigma_p \times 2 \times \boldsymbol{\Phi}_{n\times n} = \sigma_p \times 2 \times \begin{pmatrix} 0.5 & \phi_{12} & \cdots & \phi_{1n} \\ \phi_{21} & 0.5 & \cdots & \phi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & 0.5 \end{pmatrix}$$

Note that the kinship coefficient ϕ_{ij} is defined as the probability that two alleles, which are drawn at random from individual i and j, respectively, are identical-by-descent (IBD). And

given that:

$$\begin{aligned}
\boldsymbol{Y} &= (Y_1, Y_2, ..., Y_n)^T \\
\boldsymbol{X}_i &= (X_{i1}, X_{i2}, ..., X_{it}),\ \boldsymbol{X}_{n\times t} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)^T \\
\boldsymbol{G}_i &= (G_{i1}, G_{i2}, ..., G_{im}),\ \boldsymbol{G}_{n\times m} = (\boldsymbol{G}_1, ..., \boldsymbol{G}_n)^T
\end{aligned}$$

we have,

$$Y_i \mid \boldsymbol{\beta}, P_i \sim Bernoulli(\psi_i^{\boldsymbol{\beta},P_i})$$

By using the logit link function for binary traits, we can construct a linear mixed effect model for individual i as:

$$\begin{aligned}
g(\psi_i^{\boldsymbol{\beta},P_i}) = \eta_i^{\boldsymbol{\beta},P_i} &= log(\frac{\psi_i^{\boldsymbol{\beta},P_i}}{1-\psi_i^{\boldsymbol{\beta},P_i}}) \\
&= \alpha_0 + X_{i1}\alpha_1 + X_{i2}\alpha_2 + ... + X_{it}\alpha_t + G_{i1}\beta_1 + G_{i2}\beta_2 + ... + G_{im}\beta_m + P_i
\end{aligned}$$

In vector form:

$$g(\psi_i^{\boldsymbol{\beta},P_i}) = \eta_i^{\boldsymbol{\beta},P_i} = log(\frac{\psi_i^{\boldsymbol{\beta},P_i}}{1-\psi_i^{\boldsymbol{\beta},P_i}}) = \alpha_0 + \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{G}_i\boldsymbol{\beta} + P_i \tag{3.2}$$

where

$$\psi_i^{\boldsymbol{\beta},P_i} = Pr(Y_i = 1 \mid \boldsymbol{\beta}, P_i) = E[Y_i \mid \boldsymbol{\beta}, P_i] = \frac{e^{\alpha_0+\boldsymbol{X}_i\boldsymbol{\alpha}+\boldsymbol{G}_i\boldsymbol{\beta}+P_i}}{1+e^{\alpha_0+\boldsymbol{X}_i\boldsymbol{\alpha}+\boldsymbol{G}_i\boldsymbol{\beta}+P_i}} \tag{3.3}$$

In matrix form (for all n individuals):

$$g(\boldsymbol{\psi}_{n\times 1}^{\boldsymbol{\beta},\boldsymbol{P}}) = \boldsymbol{\eta}_{n\times 1}^{\boldsymbol{\beta},\boldsymbol{P}} = \alpha_0 + \boldsymbol{X}_{n\times t}\boldsymbol{\alpha}_{t\times 1} + \boldsymbol{G}_{n\times m}\boldsymbol{\beta}_{m\times 1} + \mathbb{1}_{n\times n}\boldsymbol{P}_{n\times 1} \tag{3.4}$$

where $\mathbb{1}_{n\times n}$ is an n dimension identity matrix. Thus, compared to Model (3.1) for an unrelated sample, Model (3.4) now has a random intercept $\mathbb{1}_{n\times n}\boldsymbol{P}_{n\times 1}$ to control for familial correlations in a related sample.

3.2.2 Proposed Method: Inference method

Our goal is to test whether or not rare genetic variants ($\boldsymbol{G}_{n\times m}$) are associated with the traits ($\boldsymbol{Y}_{n\times 1}$) while adjusting for the covariates ($\boldsymbol{X}_{n\times t}$) and the familial correlations ($\boldsymbol{P}_{n\times 1}$). Therefore, the null hypothesis is $H_0 : \boldsymbol{\beta}_{m\times 1} = 0$, which is equivalent to $H_0 : \boldsymbol{\tau} = 0$. Under the null hypothesis, the reduced model is:

$$g(\boldsymbol{\psi}^{\boldsymbol{P}}_{n\times 1}) = \boldsymbol{\eta}^{\boldsymbol{P}}_{n\times 1} = \alpha_0 + \boldsymbol{X}_{n\times t}\boldsymbol{\alpha}_{t\times 1} + \mathbb{1}_{n\times n}\boldsymbol{P}_{n\times 1} \tag{3.5}$$

3.2.2.1 Quasi-Likelihood In order to apply variance component score test similar to Wu et al. [2011] and Oualkacha et al. [2013] to test the null hypothesis of $H_0 : \boldsymbol{\tau} = 0$ while adjusting for covariates and familial correlation (polygenic effects) for binary traits, we derive the test statistic based on the work of Lin [1997].

First of all, we construct the log-likelihood by integrating out the random effects:

$$\begin{aligned} l(\alpha_0, \boldsymbol{\alpha}, \tau, \sigma_p) &= ln \iint L(\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{P}) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\boldsymbol{P} \\ &= ln \iint L(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{P}) \times L(\boldsymbol{\beta}) \times L(\boldsymbol{P}) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\boldsymbol{P} \\ &= ln \iint exp\left\{l(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{P})\right\} \times L(\boldsymbol{\beta}) \times L(\boldsymbol{P}) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\boldsymbol{P} \\ &= ln \iint exp\left\{l(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{P})\right\} \times L(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \times L(\boldsymbol{P}) \mathrm{d}\boldsymbol{P} \end{aligned} \tag{3.6}$$

where $l(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{P})$ is the log-likelihood function.

Since the log-likelihood function involves multiple integrals, it is very difficult to obtain an explicit form, so we apply Laplace's method to approximate the log-likelihood function

by applying Taylor Expansion at $\tau = 0$, which is equivalent to $\boldsymbol{\beta} = 0$. So,

$$\begin{aligned} exp\left\{l(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{P})\right\} &= exp\left\{\sum_{i=1}^{n} l_i(y \mid \boldsymbol{\beta}, P_i)\right\} \\ &\approx exp\left\{\sum_{i=1}^{n} l_i(y \mid 0, P_i)\right\} \times \left(1 + \sum_{i=1}^{n} \frac{\partial l_i(y \mid 0, P_i)}{\partial \eta_i^{P_i}} G_i \boldsymbol{\beta}\right. \\ &+ \frac{1}{2}\boldsymbol{\beta}^T \left[\left\{\sum_{i=1}^{n} \frac{\partial l_i(y \mid 0, P_i)}{\partial \eta_i^{P_i}} G_i^T\right\}\left\{\sum_{i=1}^{n} \frac{\partial l_i(y \mid 0, P_i)}{\partial \eta_i^{P_i}} G_i\right\}\right. \\ &\left.\left. + \sum_{i=1}^{n} \frac{\partial^2 l_i(y \mid 0, P_i)}{\partial (\eta_i^{P_i})^2} G_i^T G_i\right] \boldsymbol{\beta} + \epsilon\right) \end{aligned} \tag{3.7}$$

Then, the first layer of integral in Equation (3.6) can be approximated by taking expectation with respect to $\boldsymbol{\beta}$ and apply Taylor Expansion at $\boldsymbol{\beta} = 0$.

$$\begin{aligned} L(\boldsymbol{Y} \mid \boldsymbol{P}) &= \int exp\left\{l(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{P})\right\} \times L(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \\ &= E_{\boldsymbol{\beta}}\left[exp\left\{l(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{P})\right\}\right] \\ &= E_{\boldsymbol{\beta}}\left[exp\left\{\sum_{i=1}^{n} l_i(y \mid \boldsymbol{\beta}, P_i)\right\}\right] \\ &\approx exp\left\{\sum_{i=1}^{n} l_i(y \mid 0, P_i)\right\}\left(1 + \frac{1}{2} tr\left(\left[\sum_{i=1}^{n} \frac{\partial l_i(y \mid 0, P_i)}{\partial \eta_i^{P_i}} G_i^T \sum_{i=1}^{n} \frac{\partial l_i(y \mid 0, P_i)}{\partial \eta_i^{P_i}} G_i\right.\right.\right. \\ &\left.\left.\left. + \sum_{i=1}^{n} \frac{\partial^2 l_i(y \mid 0, P_i)}{\partial (\eta_i^{P_i})^2} G_i^T G_i\right] \boldsymbol{\Sigma}_\beta\right) + o(\beta)\right) \end{aligned} \tag{3.8}$$

Then,

$$l(\alpha_0, \boldsymbol{\alpha}, \tau, \sigma_p) = ln \int L(\boldsymbol{Y} \mid \boldsymbol{P}) \times L(\boldsymbol{P}) \mathrm{d}\boldsymbol{P} \tag{3.9}$$

Now, assume $L(\boldsymbol{Y} \mid \boldsymbol{P})$ is known from Equation (3.8), we are taking derivative with respect to τ based on Equation (3.9).

$$\begin{aligned} \frac{\partial l(\alpha_0, \boldsymbol{\alpha}, \tau, \sigma_p)}{\partial \tau} &= \frac{\partial}{\partial \tau} \int L(\boldsymbol{Y} \mid \boldsymbol{P}) \times L(\boldsymbol{P}) \mathrm{d}\boldsymbol{P} \times \frac{1}{\int L(\boldsymbol{Y} \mid \boldsymbol{P}) \times L(\boldsymbol{P}) \mathrm{d}\boldsymbol{P}} \\ &= \frac{\partial}{\partial \tau} \int L(\boldsymbol{Y} \mid \boldsymbol{P}) \times L(\boldsymbol{P}) \times \frac{1}{L(\boldsymbol{Y})} \mathrm{d}\boldsymbol{P} \end{aligned} \tag{3.10}$$

Since $L(\boldsymbol{P})$ does not contain τ, so, under regularity conditions, we can move the derivative into the integral.

$$\begin{aligned}\frac{\partial l(\alpha_0, \boldsymbol{\alpha}, \tau, \sigma_p)}{\partial \tau} &= \int \frac{\partial}{\partial \tau} L(\boldsymbol{Y} \mid \boldsymbol{P}) \times L(\boldsymbol{P}) \times \frac{1}{L(\boldsymbol{Y})} \mathrm{d}\boldsymbol{P} \\ &= \int \frac{\partial}{\partial \tau} l(\boldsymbol{Y} \mid \boldsymbol{P}) \times L(\boldsymbol{P}) \times \frac{L(\boldsymbol{Y} \mid \boldsymbol{P})}{L(\boldsymbol{Y})} \mathrm{d}\boldsymbol{P} \qquad (3.11) \\ &= \int \frac{\partial l(\boldsymbol{Y} \mid \boldsymbol{P})}{\partial \tau} \times L(\boldsymbol{P} \mid \boldsymbol{Y}) \mathrm{d}\boldsymbol{P} \\ &= E_{\boldsymbol{P}} \left[\frac{\partial l(\boldsymbol{Y} \mid \boldsymbol{P})}{\partial \tau} \mid \boldsymbol{Y} \right] \qquad (3.12)\end{aligned}$$

Note that we have already derived the likelihood in Equation (3.8), according to Lin [1997], we can take the logarithm and approximate it.

$$\begin{aligned}l(\boldsymbol{Y} \mid \boldsymbol{P}) &= log(L(\boldsymbol{Y} \mid \boldsymbol{P})) \\ &= \sum_{i=1}^{n} l_i(y \mid 0, P_i) + log\Bigg[1 + \frac{1}{2} tr\Bigg(\Bigg[\sum_{i=1}^{n} \frac{\partial l_i(y \mid 0, P_i)}{\partial \eta_i^{P_i}} G_i^T \sum_{i=1}^{n} \frac{\partial l_i(y \mid 0, P_i)}{\partial \eta_i^{P_i}} G_i \\ &+ \sum_{i=1}^{n} \frac{\partial^2 l_i(y \mid 0, P_i)}{\partial (\eta_i^{P_i})^2} G_i^T G_i \Bigg] \boldsymbol{\Sigma}_\beta \Bigg) + o(\beta) \Bigg] \qquad (3.13)\end{aligned}$$

Note that, by applying first order Taylor expansion at 0 to the 'log' part in Equation (3.13), similar to

$$\begin{aligned}log(1 + x) &\approx log(1 + 0) + \frac{1}{1 + 0} \times (x - 0) \\ &= x \qquad (3.14)\end{aligned}$$

we can get

$$\begin{aligned}l(\boldsymbol{Y} \mid \boldsymbol{P}) &= log(L(\boldsymbol{Y} \mid \boldsymbol{P})) \\ &= \sum_{i=1}^{n} l_i(y \mid 0, P_i) + \frac{1}{2} tr\Bigg(\Bigg[\sum_{i=1}^{n} \frac{\partial l_i(y \mid 0, P_i)}{\partial \eta_i^{P_i}} G_i^T \sum_{i=1}^{n} \frac{\partial l_i(y \mid 0, P_i)}{\partial \eta_i^{P_i}} G_i \\ &+ \sum_{i=1}^{n} \frac{\partial^2 l_i(y \mid 0, P_i)}{\partial (\eta_i^{P_i})^2} G_i^T G_i \Bigg] \boldsymbol{\Sigma}_\beta \Bigg) + o(\beta) \qquad (3.15)\end{aligned}$$

Now, according to Equation (3.12), we take derivative with respect to τ based on Equation (3.15), and in matrix form, we can get

$$\frac{\partial l(\boldsymbol{Y} \mid \boldsymbol{P})}{\partial \tau} = \frac{1}{2} tr\left(\boldsymbol{G}^T \left[\frac{\partial l(\boldsymbol{Y} \mid \boldsymbol{0}, \boldsymbol{P})}{\partial \boldsymbol{\eta^P}} \frac{\partial l(\boldsymbol{Y} \mid \boldsymbol{0}, \boldsymbol{P})}{(\boldsymbol{\eta^P})^T} + \frac{\partial^2 l(\boldsymbol{Y} \mid \boldsymbol{0}, \boldsymbol{P})}{\partial \boldsymbol{\eta^P} (\boldsymbol{\eta^P})^T} \right] \boldsymbol{G} \boldsymbol{W}_{\boldsymbol{m} \times \boldsymbol{m}} \right) \tag{3.16}$$

Then, according to the properties of quasi-likelihood from Wedderburn [1974],

$$\begin{aligned} \frac{\partial l(\boldsymbol{Y} \mid \boldsymbol{0}, \boldsymbol{P})}{\partial \boldsymbol{\eta^P}} &= \frac{\partial l(\boldsymbol{Y} \mid \boldsymbol{0}, \boldsymbol{P})}{\partial \boldsymbol{\psi^P}} \frac{\partial \boldsymbol{\psi^P}}{\partial \boldsymbol{\eta^P}} \\ &= \frac{\boldsymbol{Y} - \boldsymbol{\psi^P}}{V(\boldsymbol{\psi^P})} \frac{1}{g'(\boldsymbol{\psi^P})} \end{aligned} \tag{3.17}$$

is an $n \times 1$ vector with the i^{th} element $\frac{y_i - \psi_i^{P_i}}{V(\psi_i^{P_i})} \frac{1}{g'(\psi_i^{P_i})}$ and $\boldsymbol{\eta^P} = g(\boldsymbol{\psi^P})$ in model (3.5), and

$$\begin{aligned} V(\psi_i^{P_i}) &= var(y_i \mid 0, P_i) = \psi_i^{P_i}(1 - \psi_i^{P_i}) \\ g(\psi_i^{P_i}) &= log(\frac{\psi_i^{P_i}}{1 - \psi_i^{P_i}}) \\ g'(\psi_i^{P_i}) &= \frac{1}{\psi_i^{P_i}(1 - \psi_i^{P_i})} \end{aligned} \tag{3.18}$$

for binary traits. And, $\frac{\partial^2 l(\boldsymbol{Y} | \boldsymbol{0}, \boldsymbol{P})}{\partial \boldsymbol{\eta^P} (\boldsymbol{\eta^P})^T}$ is an $n \times n$ diagonal matrix with the elements $\frac{\partial^2 l(y_i | 0, P_i)}{\partial (\eta_i^{P_i})^2}$ on

the diagonal, where

$$\begin{aligned}
\frac{\partial^2 l(y_i \mid 0, P_i)}{\partial (\eta_i^{P_i})^2} &= \frac{\partial}{\partial \eta_i^{P_i}} \left(\frac{\partial l(y_i \mid 0, P_i)}{\partial \eta_i^{P_i}} \right) \\
&= \frac{\partial}{\partial \eta_i^{P_i}} \left(\frac{\partial l(y_i \mid 0, P_i)}{\partial \psi_i^{P_i}} \frac{\partial \psi_i^{P_i}}{\partial \eta_i^{P_i}} \right) \\
&= \frac{\partial}{\partial \eta_i^{P_i}} \left(\frac{\partial l(y_i \mid 0, P_i)}{\partial \psi_i^{P_i}} \right) \frac{\partial \psi_i^{P_i}}{\partial \eta_i^{P_i}} + \frac{\partial l(y_i \mid 0, P_i)}{\partial \psi_i^{P_i}} \frac{\partial}{\partial \eta_i^{P_i}} \left(\frac{\partial \psi_i^{P_i}}{\partial \eta_i^{P_i}} \right) \\
&= \frac{\partial}{\partial \eta_i^{P_i}} \left(\frac{y_i - \psi_i^{P_i}}{V(\psi_i^{P_i})} \right) \frac{1}{g^{'}(\psi_i^{P_i})} + \frac{y_i - \psi_i^{P_i}}{V(\psi_i^{P_i})} \frac{\partial}{\partial \eta_i^{P_i}} \left(\frac{1}{g^{'}(\psi_i^{P_i})} \right) \\
&= \left[\frac{\partial (y_i - \psi_i^{P_i})}{\partial \eta^{P_i}} \frac{1}{V(\psi_i^{P_i})} + (y_i - \psi_i^{P_i}) \frac{\partial}{\partial \eta_i^{P_i}} (\frac{1}{V(\psi_i^{P_i})}) \right] \frac{1}{g^{'}(\psi_i^{P_i})} \\
&+ \frac{y_i - \psi_i^{P_i}}{V(\psi_i^{P_i})} \frac{\partial (g^{'}(\psi_i^{P_i}))^{-1}}{\partial \eta_i^{P_i}} \\
&= \left[-\frac{1}{g^{'}(\psi_i^{P_i})\ V(\psi_i^{P_i})} - (y_i - \psi_i^{P_i}) [V(\psi_i^{P_i})]^{-2} \frac{V^{'}(\psi_i^{P_i})}{g^{'}(\psi_i^{P_i})} \right] \frac{1}{g^{'}(\psi_i^{P_i})} \\
&- \frac{y_i - \psi_i^{P_i}}{V(\psi_i^{P_i})} [g^{'}(\psi_i^{P_i})]^{-2} \frac{g^{''}(\psi_i^{P_i})}{g^{'}(\psi_i^{P_i})} \\
&= -\left[\frac{1}{V(\psi_i^{P_i})[g^{'}(\psi_i^{P_i})]^2} + \frac{V^{'}(\psi_i^{P_i}) g^{'}(\psi_i^{P_i}) + V(\psi_i^{P_i}) g^{''}(\psi_i^{P_i})}{[V(\psi_i^{P_i})]^2 [g^{'}(\psi_i^{P_i})]^3} (y_i - \psi_i^{P_i}) \right] \qquad (3.19)
\end{aligned}$$

in which $V^{'}(\psi_i^{P_i}) = 1 - 2\psi_i^{P_i}$ and $g^{''}(\psi_i^{P_i}) = \frac{2\psi_i^{P_i} - 1}{(\psi_i^{P_i}(1 - \psi_i^{P_i}))^2}$. Note that, in Equation (3.19), for binary traits y_i,

$$\begin{aligned}
V^{'}(\psi_i^{P_i}) g^{'}(\psi_i^{P_i}) + V(\psi_i^{P_i}) g^{''}(\psi_i^{P_i}) &= \frac{1 - 2\psi_i^{P_i}}{\psi_i^{P_i}(1 - \psi_i^{P_i})} + \frac{\psi_i^{P_i}(1 - \psi_i^{P_i})(2\psi_i^{P_i} - 1)}{(\psi_i^{P_i}(1 - \psi_i^{P_i}))^2} \\
&= 0 \qquad (3.20)
\end{aligned}$$

Therefore,

$$\frac{\partial^2 l(y_i \mid 0, P_i)}{\partial (\eta_i^{P_i})^2} = -\frac{1}{V(\psi_i^{P_i})[g^{'}(\psi_i^{P_i})]^2} \qquad (3.21)$$

Let's set

$$\boldsymbol{\Omega}_{n \times n}^{P} = -\frac{\partial^2 l(\boldsymbol{Y} \mid \boldsymbol{0}, \boldsymbol{P})}{\partial \boldsymbol{\eta}^{\boldsymbol{P}} (\boldsymbol{\eta}^{\boldsymbol{P}})^T} = diag[-\frac{\partial^2 l(y_i \mid 0, P_i)}{\partial (\eta_i^{P_i})^2}] = diag[\frac{1}{V(\psi_i^{P_i})[g^{'}(\psi_i^{P_i})]^2}] \qquad (3.22)$$

$$\boldsymbol{\Delta}_{n \times n} = diag[\frac{1}{g^{'}(\psi_i^{P_i})}] \qquad (3.23)$$

Then, Equation (3.16) becomes

$$\begin{aligned}\frac{\partial l(\boldsymbol{Y} \mid \boldsymbol{P})}{\partial \tau} &= \frac{1}{2} tr\left(\boldsymbol{G}^T\left[\boldsymbol{\Omega}^P \boldsymbol{\Delta}^{-1}(\boldsymbol{Y}-\boldsymbol{\psi}^{\boldsymbol{P}})(\boldsymbol{Y}-\boldsymbol{\psi}^{\boldsymbol{P}})^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Omega}^P - \boldsymbol{\Omega}^P\right]\boldsymbol{G}\boldsymbol{W}_{m\times m}\right) \\ &= \frac{1}{2}\left((\boldsymbol{Y}-\boldsymbol{\psi}^{\boldsymbol{P}})^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Omega}^P\boldsymbol{G}\boldsymbol{W}_{m\times m}\boldsymbol{G}^T\boldsymbol{\Omega}^P\boldsymbol{\Delta}^{-1}(\boldsymbol{Y}-\boldsymbol{\psi}^{\boldsymbol{P}}) - tr(\boldsymbol{G}^T\boldsymbol{\Omega}^P\boldsymbol{G}\boldsymbol{W}_{m\times m})\right)\end{aligned} \tag{3.24}$$

Therefore, if we put Equation (3.24) back into Equation (3.12), we can get

$$\begin{aligned}&\frac{\partial l(\alpha_0, \boldsymbol{\alpha}, \tau, \sigma_p)}{\partial \tau} = E_{\boldsymbol{P}}\left[\frac{\partial l(\boldsymbol{Y} \mid \boldsymbol{P})}{\partial \tau} \mid \boldsymbol{Y}\right] \\ &= \frac{1}{2}E_{\boldsymbol{P}}\left[\left((\boldsymbol{Y}-\boldsymbol{\psi}^{\boldsymbol{P}})^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Omega}^P\boldsymbol{G}\boldsymbol{W}_{m\times m}\boldsymbol{G}^T\boldsymbol{\Omega}^P\boldsymbol{\Delta}^{-1}(\boldsymbol{Y}-\boldsymbol{\psi}^{\boldsymbol{P}}) - tr(\boldsymbol{G}^T\boldsymbol{\Omega}^P\boldsymbol{G}\boldsymbol{W}_{m\times m})\right) \mid \boldsymbol{Y}\right]\end{aligned} \tag{3.25}$$

3.2.2.2 Score function From the log-likelihood, we can take derivatives with respect to $\boldsymbol{\beta}$, after some calculations, according to Lin [1997] and Zhang and Lin [2003], the score function is in the following form:

$$\begin{aligned}\boldsymbol{U}^*_{\boldsymbol{\beta}}(\hat{\alpha_0}, \hat{\boldsymbol{\alpha}}, \widehat{\sigma_p}) &= \frac{1}{2}(\boldsymbol{Y}^* - \boldsymbol{X}\boldsymbol{\alpha})^T\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{G}_{n\times m}\boldsymbol{W}_{m\times m}\boldsymbol{G}^T_{n\times m}\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{Y}^* - \boldsymbol{X}\boldsymbol{\alpha}) \\ &\quad - \frac{1}{2}tr(\boldsymbol{W}^T_{m\times m}\boldsymbol{G}^T_{n\times m}\boldsymbol{\Lambda}_\tau\boldsymbol{G}_{n\times m}) \\ &= \boldsymbol{U} - \boldsymbol{e}\end{aligned} \tag{3.26}$$

where $\boldsymbol{U}$ and $\boldsymbol{e}$ are the first and second component in Equation 3.26 respectively. And

$$\boldsymbol{V}_{\boldsymbol{\beta}} = (\boldsymbol{\Omega}^P_{n\times n})^{-1} + \mathbb{1}_{n\times n}\boldsymbol{\Sigma}_P\mathbb{1}^T_{n\times n} \tag{3.27}$$

$$\boldsymbol{\Sigma}_P = \widehat{\sigma_p} \times 2 \times \boldsymbol{\Phi}_{n\times n} \tag{3.28}$$

$$\boldsymbol{\Lambda}_\tau = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} - \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}_{\boldsymbol{\beta}}^{-1} \tag{3.29}$$

and

$$\boldsymbol{\Omega}^P_{n\times n} = E(diag([V(\psi_i^{P_i})g'(\psi_i^{P_i})^2]^{-1})) \tag{3.30}$$

where

$$V(\psi_i^{P_i}) = var(y_i \mid P_i) = \psi_i^{P_i}(1 - \psi_i^{P_i}) \tag{3.31}$$

$$\psi_i^{P_i} = Pr(Y_i = 1 \mid P_i) = E[Y_i \mid P_i] = \frac{e^{\alpha_0 + \boldsymbol{X}_i\boldsymbol{\alpha} + P_i}}{1 + e^{\alpha_0 + \boldsymbol{X}_i\boldsymbol{\alpha} + P_i}} \tag{3.32}$$

Note that $g(\psi_i^{P_i})$ is the i^{th} element in Equation (3.5) and $(\alpha_0, \boldsymbol{\alpha})$ in above equations are evaluated at $(\hat{\alpha_0}, \hat{\boldsymbol{\alpha}})$. In order to calculate $\boldsymbol{V_\beta}$, one need to obtain the accurate estimates of $(\sigma_p, \psi_i^{P_i})$; however, since these estimates are difficult to obtain, we have applied the R function "glmmPQL" [Schall, 1991; Breslow and Clayton, 1993; Wolfinger and O'connell, 1993] in the R package "MASS" [Venables and Ripley, 2002] to fit the model (3.5), and we have directly obtained the estimate of $\boldsymbol{V_\beta}$ by using the R function "extract.lme.cov", thus, $\boldsymbol{U}^*_{\boldsymbol{\beta}}(\hat{\alpha_0}, \hat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\sigma}_p})$ can be calculated.

The R function "glmmPQL" repeatedly calls the R function "lme" [Pinheiro et al., 2013] to fit the model until the estimates of parameters are close enough (user specified tolerance). $\boldsymbol{Y}^*$ is the value of Y at convergence, which is estimated by assuming the working vector $\boldsymbol{Y}^* = \boldsymbol{X\alpha} + \boldsymbol{P} + \boldsymbol{\Delta}^{-1}(\boldsymbol{Y} - \boldsymbol{\psi^P})$ during the iteration process.

3.2.2.3 Information matrix Given the expression in Equation (3.27), if we take derivative of $\boldsymbol{V}_\beta$ with respect to σ_p and let $\boldsymbol{K} = 2 \times \boldsymbol{\Phi}$, then we have $\frac{\partial \boldsymbol{V_\beta}}{\partial \sigma_p} = 2 \times \boldsymbol{\Phi} = \boldsymbol{K}$. According to Zhang and Lin [2003], the conditional information matrix given nuisance parameters for testing $H_0 : \tau = 0$ is

$$\boldsymbol{I}_{\tau|\sigma_p} = \boldsymbol{I}_{\tau\tau} - \boldsymbol{I}_{\tau\sigma_p}\boldsymbol{I}^{-1}_{\sigma_p\sigma_p}\boldsymbol{I}_{\sigma_p\tau} \tag{3.33}$$

where

$$\boldsymbol{I}_{\tau\tau} = \frac{1}{2}tr(\boldsymbol{\Lambda}_\tau \boldsymbol{GWG}^T \boldsymbol{\Lambda}_\tau \boldsymbol{GWG}^T) \tag{3.34}$$

$$\boldsymbol{I}_{\tau\sigma_p} = \frac{1}{2}tr(\boldsymbol{\Lambda}_\tau \boldsymbol{GWG}^T \boldsymbol{\Lambda}_\tau \boldsymbol{K}) \tag{3.35}$$

$$\boldsymbol{I}_{\sigma_p\sigma_p} = \frac{1}{2}tr(\boldsymbol{\Lambda}_\tau \boldsymbol{K} \boldsymbol{\Lambda}_\tau \boldsymbol{K}) \tag{3.36}$$

3.2.2.4 Q-statistic According to Zhang and Lin [2003], Zhang and Lin [2008] and Huang and Zhang [2008], when the parameter is tested at the boundary of its domain, for example, in our case, variance component τ has domain $[0, \infty]$ and it is tested at $\tau = 0$, the asymptotic distribution of $\frac{\boldsymbol{U}^*_{\boldsymbol{\beta}}(\hat{\alpha_0},\hat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\sigma}_p})^2}{\boldsymbol{I}_{\tau|\sigma_p}}$ does not follow a chi-square distribution with one degree of freedom. Rather, the null distribution of $\boldsymbol{U}$ in Equation (3.26) can be approximated by a scaled chi-square distribution. Therefore, we have used a scaled chi-square distribution for our test statistic. Zhang and Lin [2003] has provided a method to calculate the scale parameter k and the corresponding degree of freedom v:

$$k = \frac{\boldsymbol{I}_{\tau|\sigma_p}}{2 \times \boldsymbol{e}} \tag{3.37}$$

$$v = \frac{2 \times \boldsymbol{e}^2}{\boldsymbol{I}_{\tau|\sigma_p}} \tag{3.38}$$

where $\boldsymbol{e}$ is the second component in Equation (3.26). Now we have our variance component score test Q as:

$$\boldsymbol{Q}\text{-}test = \frac{\boldsymbol{U}}{k} \tag{3.39}$$

which, under the null hypothesis, asymptotically follows a chi-square distribution with v degree of freedom. We have implemented this test statistic in an R function 'Qtest' (Appendix C) .

3.3 SIMULATION

In order to evaluate Type I error and power of our statistic (Qtest), we applied it on simulated data and compared it to six other statistics. First, we simulated three-generation family structures according to Figure 3.1, in which the number of offspring in each sub-family was generated from a negative binomial distribution with dispersion parameter 2.84 and probability 0.93. Note that the family structures varied from family to family within each dataset, but were kept the same from dataset to dataset for all simulated scenarios in order to obtain

consistent sample sizes. Please see Appendix B for all the 25 simulated families. Then, we used a haplotype data pool that was generated by the calibrated coalescent model [Schaffner et al., 2005] with mimicking the linkage disequilibrium (LD) structure of European ancestry. This haplotype pool contained 10,000 haplotypes, and covered 200 kb region on chromosome

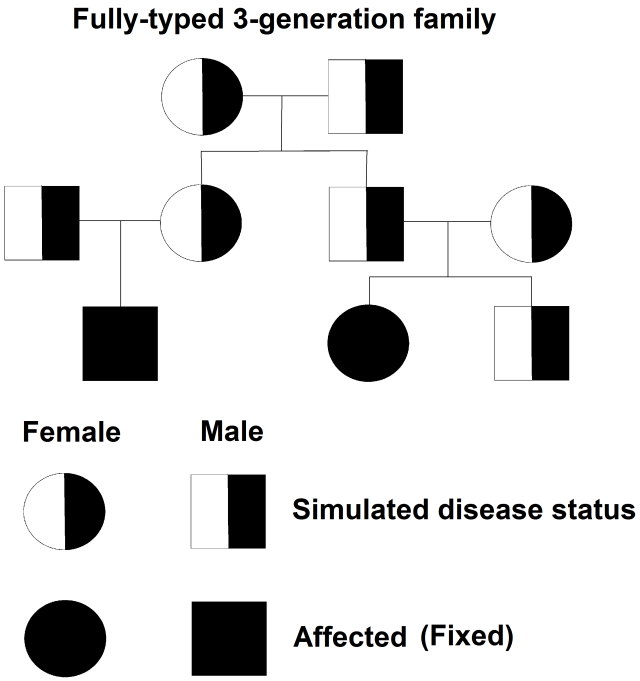


Figure 3.1: An example of simulated family structure. The number of offspring are randomly generated from a negative binomial distribution. Please see Appendix B for all the 25 simulated families.

one. We selected rare variants between positions 79 and 5,427, in which there were 100 polymorphic markers with MAF smaller than 0.05. To simulate genotypes, we first randomly chose haplotypes from the pool and assigned them to the founders. Second, we assigned haplotypes to the other individuals by mimicking a gene-dropping process. Specifically, for each individual who has parents in the data, assuming no recombination, we randomly chose one haplotype from his/her father and another one from his/her mother as his/her two haplotypes, respectively. Finally, we calculated genotypes (coded as 0, 1, 2) from the assigned haplotypes.

In order to generate traits, within each scenario, we first randomly selected different percentages of markers to be risk or protective, and then assigned different odds ratios (OR), fixed or MAF-dependent (O^+ for risk; O^- for protective) [Wu et al., 2011], to the markers to construct different scenarios in Table 3.2, in which risk variants have OR > 1 and protective variants have OR < 1. Note that the MAF was calculated from the larger haplotype pool instead of the much smaller sampled dataset. Then, we used the logistic model below to generate the probability of being affected ($Prob(Y_i = 1)$) for the ith individual.

$$Logit(Prob(Y_i = 1)) = b_0 + \sum_{j=1}^{m} ln(OR_j) \times g_{ij} + P_i; \ (i = 1 \cdots n, \ j = 1 \cdots m) \tag{3.40}$$

where b_0 is calculated from the prevalence, which we set at 5%; j represented the jth marker, m was the total number of markers, n was the sample size. OR_j represented the odds ratio for marker j. And P_i was the polygenic effect generated from a multivariate normal distribution below.

$$\boldsymbol{P} = (P_1, P_2, ..., P_n)^T \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_P) \tag{3.41}$$

$$\text{where } \boldsymbol{\Sigma}_P = \sigma_p \times 2 \times \boldsymbol{\Phi}_{n\times n} = \sigma_p \times 2 \times \begin{pmatrix} 0.5 & \phi_{12} & \cdots & \phi_{1n} \\ \phi_{21} & 0.5 & \cdots & \phi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & 0.5 \end{pmatrix}$$

And ϕ_{ik} was expected kinship coefficient between subjects i and k. Thus, $\boldsymbol{\Phi}_{n\times n}$ could be calculated directly from the family structure by using the R function 'kinship' from the R-package 'kinship2' [Therneau et al., 2014]. For example, $\boldsymbol{\Phi}_{n\times n}$ for the family in Figure 3.1 was in the form below:

$$\boldsymbol{\Phi}_{9\times 9} = \begin{pmatrix} 0.50 & 0.00 & 0.25 & 0.25 & 0.00 & 0.00 & 0.12 & 0.12 & 0.12 \\ 0.00 & 0.50 & 0.25 & 0.25 & 0.00 & 0.00 & 0.12 & 0.12 & 0.12 \\ 0.25 & 0.25 & 0.50 & 0.25 & 0.00 & 0.00 & 0.25 & 0.12 & 0.12 \\ 0.25 & 0.25 & 0.25 & 0.50 & 0.00 & 0.00 & 0.12 & 0.25 & 0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.00 & 0.00 & 0.25 & 0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.25 & 0.00 & 0.00 \\ 0.12 & 0.12 & 0.25 & 0.12 & 0.00 & 0.25 & 0.50 & 0.06 & 0.06 \\ 0.12 & 0.12 & 0.12 & 0.25 & 0.25 & 0.00 & 0.06 & 0.50 & 0.25 \\ 0.12 & 0.12 & 0.12 & 0.25 & 0.25 & 0.00 & 0.06 & 0.25 & 0.50 \end{pmatrix}$$

Here we set σ_p at 0.38, which was calculated from the variance of $\sum_{j=1}^{m} ln(OR_j) \times g_{ij}, i = 1 \cdots n$, over all n subjects. The reason why we set σ_p at 0.38 was because we did not want the polygenic effect to overwhelm the rare variant effect, thus we restricted the variance of the polygenic effect to be equal to the variance of rare variant effects. To simulate the data under null hypothesis, after generating traits by using Equation 3.40, we generated another set of genotypes. This way, the number and pattern of affected individuals was kept fixed, but there was no association between genotypes and traits. In this manner, for each replication, we ascertained the same set of 25 families that contained at least two affected subjects in the youngest generation, sample size is 633. We only kept the markers whose MAF = (0, 0.05]. Table 3.2 shows the average number of markers that were analyzed in the data. For each scenario, we simulated 3,000 datasets to measure Type I error, and 3,000 datasets to measure power.

We compared Qtest to six other statistics. First, the FSKAT test developed by Yan et al. [2015] is mathematically very similar to Qtest when estimating the parameters. The main differences were that, in FSKAT, the test statistic was constructed mainly by using Penalized Quasi-likelihood, and followed a mixture of chi-square distributions. In specific, FSKAT constructed a Q-statistic where

$$\boldsymbol{Q} = (\boldsymbol{Y}^* - \boldsymbol{X\alpha})^T \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} \boldsymbol{G}_{n\times m} \boldsymbol{W}_{m\times m} \boldsymbol{G}_{n\times m}^T \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{Y}^* - \boldsymbol{X\alpha}) \tag{3.42}$$

Q follows a mixture of chi-square distributions under the null hypothesis.

$$\boldsymbol{Q} \sim \sum_{j=1}^{m} \lambda_j \chi^2_{1,j} \tag{3.43}$$

where $\chi^2_{1,j}$ represented a chi-square distribution with one degree of freedom, λ_j were the eigenvalues of the matrix $\boldsymbol{W}_{m\times m}\boldsymbol{G}^T_{n\times m}\boldsymbol{V}^{-1}_{\boldsymbol{\beta}}\boldsymbol{P}_0\boldsymbol{V}^{-1}_{\boldsymbol{\beta}}\boldsymbol{G}_{n\times m}\boldsymbol{W}_{m\times m}$, and $\boldsymbol{P}_0$ was the variance of $(\boldsymbol{Y}^* - \boldsymbol{X}\boldsymbol{\alpha})$. While, as in Equation 3.39, Qtest was constructed mainly by using the Laplace method, and follows a scaled chi-square distribution.

Second, we also compared Qtest to two statistics developed by Schaid et al. [2013], which are the Burden and Kernel statistics (R-package: 'pedgene') that treat the traits as fixed, genotypes as random, and carry out burden and kernel test statistics to identify the association. By treating the traits as fixed, the covariance among markers can be calculated as

$$Cov(G_j, G_k) = w_j w_k \sum_{i=1}^{n}\sum_{l=1}^{n}(y_i - \hat{y_i})(y_l - \hat{y_l})Cov(g_{ij}, g_{lk}) \tag{3.44}$$

where G_j and G_k are two vectors contain marker genotypes for marker j and marker k, respectively, over all subjects in the sample. Therefore, specifically, the Burden test is:

$$T = \frac{\left[(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T\boldsymbol{S}\right]^2}{(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T\boldsymbol{V}_S(\boldsymbol{Y} - \hat{\boldsymbol{Y}})} \tag{3.45}$$

where $\boldsymbol{V}_S$ is a function of $Cov(G_j, G_k)$ and the kinship matrix $\boldsymbol{\Phi}_{n\times n}$, and $\boldsymbol{S} = (S_1, S_2, ..., S_n)$ for all the n subjects in the sample and $S_i = \sum_{j=1}^{m} w_j g_{ij}$ where w_j is the weight for marker j and g_{ij} is the genotype of marker j for subject i. And T has an approximate chi-square distribution with one degree of freedom. For the Kernel test, it constructs a Q function:

$$\boldsymbol{Q} = \sum_{j=1}^{m}\left[w_j\sum_{i=1}^{n}(y_i - \hat{y_i})g_{ij}\right]^2 \tag{3.46}$$

Similar to the Qtest, this Q statistic also follows a mixture of independent chi-square distributions, and is approximated by a scaled chi-square distribution, in which the scale parameter can be calculated as $Var(Q)/(2E(Q))$ and the degrees of freedom of the scaled chi-square

distribution can be calculated as $2(E(Q))^2/Var(Q)$, where $E(Q)$ and $Var(Q)$ are functions of $Cov(G_j, G_k)$ and the kinship matrix $\mathbf{\Phi}_{n\times n}$.

Third, we compared Qtest to famSKAT [Chen et al., 2013] and FFBSKAT [Svishcheva et al., 2014], which have been compared in Svishcheva et al. [2014]. Note that, these two statistics are designed for quantitative traits, here we apply them such that we treat a binary trait as a continuous one. These two methods were the fast implementations of the methods proposed by Schifano et al. [2012]; Chen et al. [2013]; Oualkacha et al. [2013] and Svishcheva et al. [2014] has showed that FFBSKAT is faster and has more features than famSKAT such as using genomic kinship matrix. Specifically, these methods used an efficient kernel machine-based regression approach to identify the association between rare genetic variants and continuous traits on family data. Finally, we also applied the GEE-based method (R function "score_FSKAT_IC_pertu" in the package "gskat") developed by Wang et al. [2013] that allowed for mis-specification of family structure. And we have obtained the p-value calculated by using 'Rademacher' perturbation adjustment method for small sample size [Wang et al., 2013]. Wang et al. [2013] applied a GEE method to quantitative traits, but they did not apply it to binary traits. Thus, in our study, we applied this method to binary traits and evaluated its performance. All simulation and comparison were done in R.

Note that we applied different weighting methods for these statistics. In specific, for Schaid's methods (Burden, Kernel), we applied equal (E) weight ($\boldsymbol{W}_{m\times m} = \mathbb{1}_{m\times m}$), sample-MAF-dependent (M) weight, which were generated from a beta distribution, Beta(MAF_j, a = 1, b = 25), where MAF_j is the minor allele frequency calculated based on the sampled dataset for the jth marker, and Madsen-Browning weight [Madsen and Browning, 2009]. For other methods except for GEE, we applied equal (E) and sample-MAF-dependent (M) weight. For GEE, we only applied sample-MAF-dependent (M) weight because in its R implementation, it was hard to set the weights to be equal.

3.4 RESULTS

As shown in Table 3.2, we have simulated eight different scenarios by setting different percentages of and assigning different odds ratios (OR) to risk (OR $>$ 1) and protective (OR $<$ 1) rare variants. The average numbers of affected subjects across all scenarios ranged from 79 to 94. The average number of variants by minor allele frequency (MAF) are about the same respectively for each frequency range across eight different scenarios.

Table 3.2: Eight simulated scenarios

	Odds Ratio	Percentage	Average number of	Average number of variants by MAF		
Scenarios	(Risk/Protective)	(Risk/Protective)	affected subjects	(0, 0.001]	(0.001, 0.01]	(0.01, 0.05]
1	1.5 / 0.5	60 / 20	78.18	7.28	10.32	5.94
2	1.5 / 1	60 / 0	80.74	7.53	10.52	5.96
3	2.5 / 0.5	60 / 20	87.27	7.73	10.69	5.96
4	2.5 / 1	60 / 0	94.01	8.02	10.80	5.96
5	O^+ / O^-	30 / 20	81.29	8.27	10.92	5.96
6	O^+ / 1	30 / 0	88.29	8.42	10.92	5.96
7	O^+ / O^-	40 / 20	93.47	7.99	10.58	5.96
8	O^+ / 1	40 / 0	89.57	8.99	10.93	5.96

MAF-dependent odds ratio: $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$

MAF_j: minor allele frequency for the jth marker in the overall haplotype pool.

3.4.1 Type I error

The Type I errors for all statistics under eight different trait simulation scenarios and three different alpha levels are summarized in Figures 3.2 , 3.3 and 3.4. We have calculated 95% confidence intervals (C.I.) for all the three alpha levels based on 10,000 replicates:

C.I. for alpha level 0.05 $= 0.05 \pm 1.96\times\sqrt{(0.05 \times (1-0.05)/10000)} = [0.046, 0.054]$

C.I. for alpha level 0.01 $= 0.01 \pm 1.96\times\sqrt{(0.01 \times (1-0.01)/10000)} = [0.008, 0.012]$

C.I. for alpha level 0.001 $= 0.001 \pm 1.96\times\sqrt{(0.001 \times (1-0.001)/10000)} = [0.0004, 0.0016]$

Note that, in each scenario, we first simulate the trait by using one set of markers with percentages and odds ratios set according to the scenario settings, and then we simulate another set of null markers independent of the trait and apply statistics to test for association between the new set of markers and the traits. In this manner, we are trying to simulate the clustering of traits within an ascertained sample under the null hypothesis.

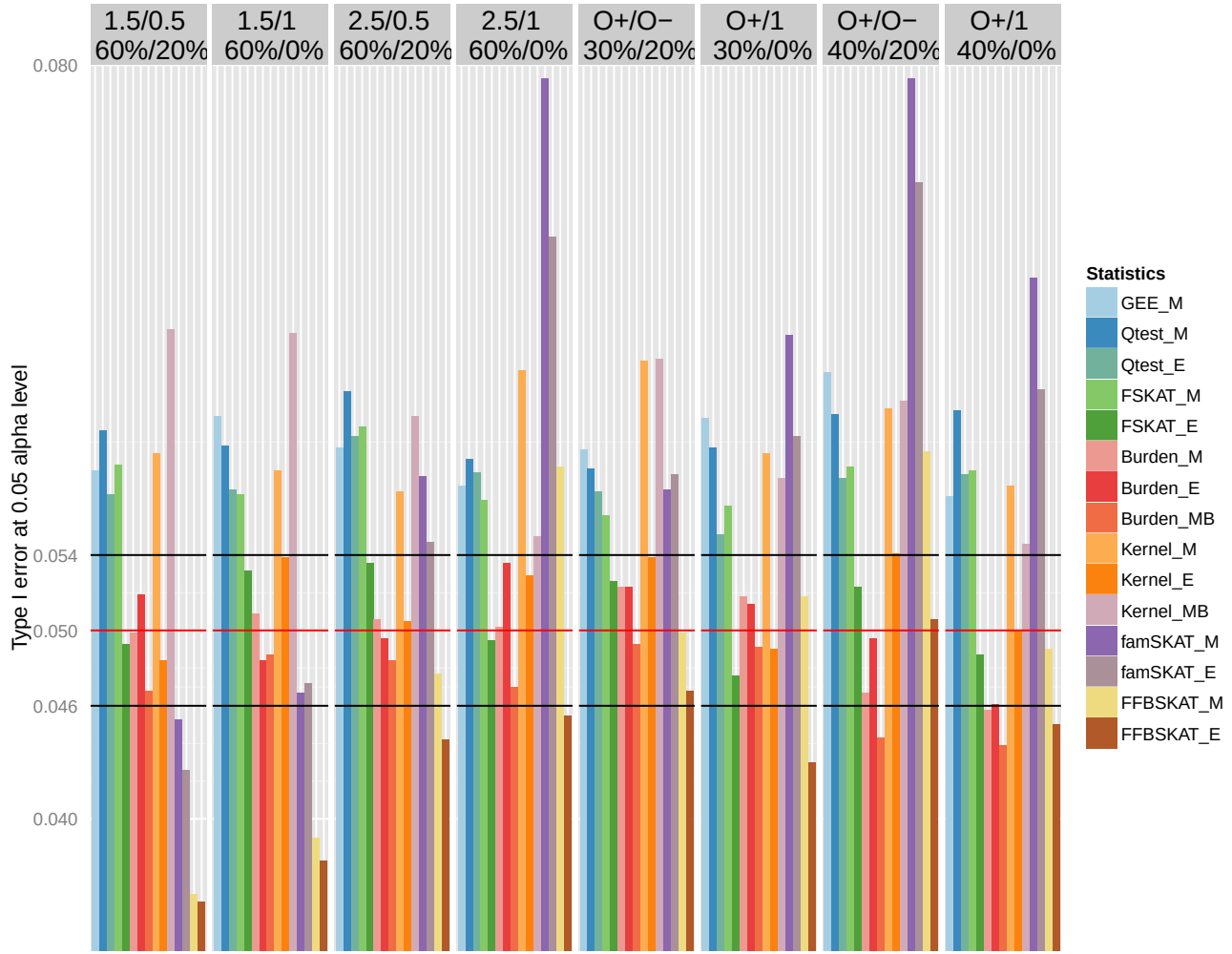


Figure 3.2: Type I error at a gene independent of the trait locus under eight trait simulation scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.05 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. The boundaries of the 95% Confidence Interval are marked out with two black lines. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights.

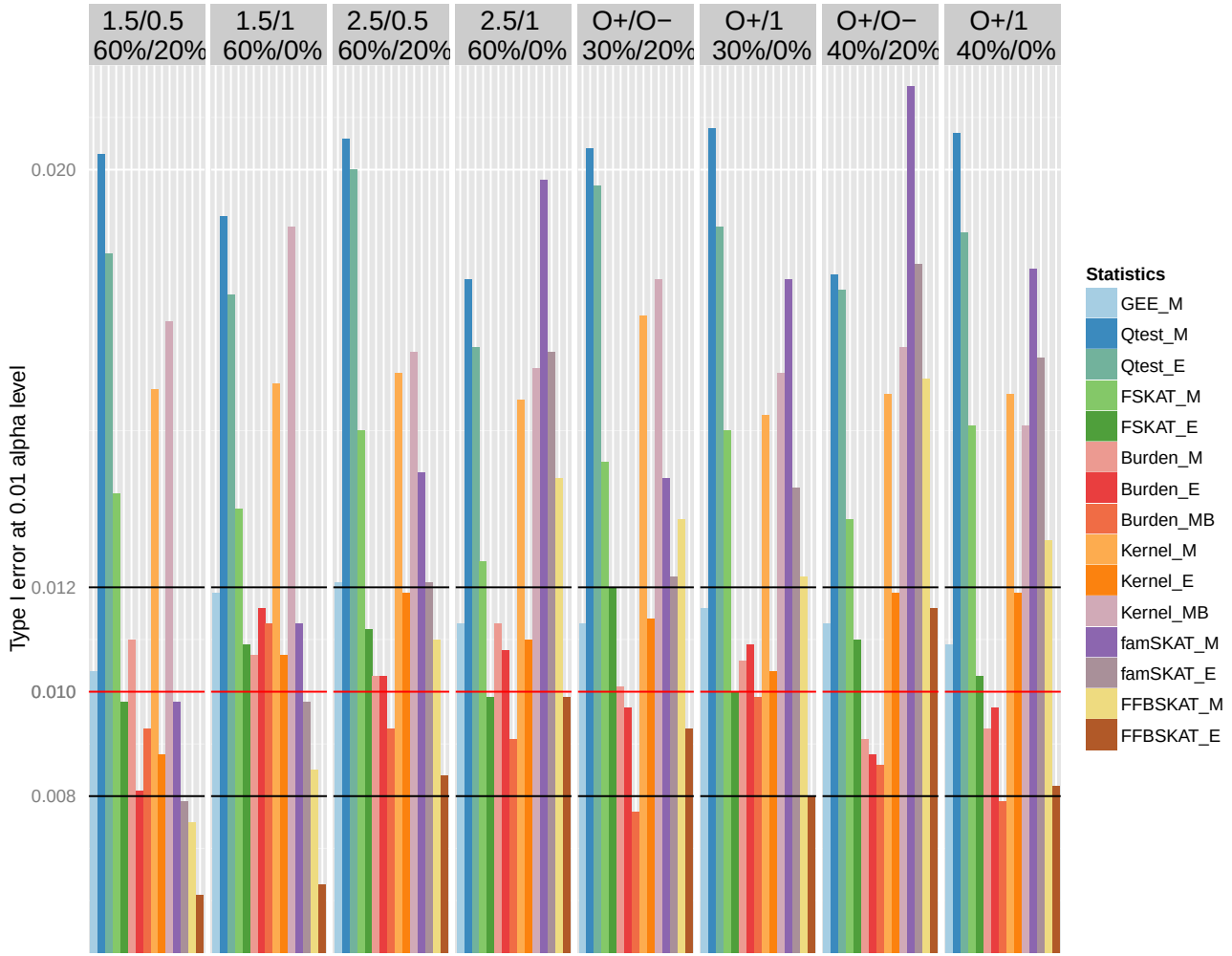


Figure 3.3: Type I error at a gene independent of the trait locus under eight trait simulation scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.01 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. The boundaries of the 95% Confidence Interval are marked out with two black lines. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights.

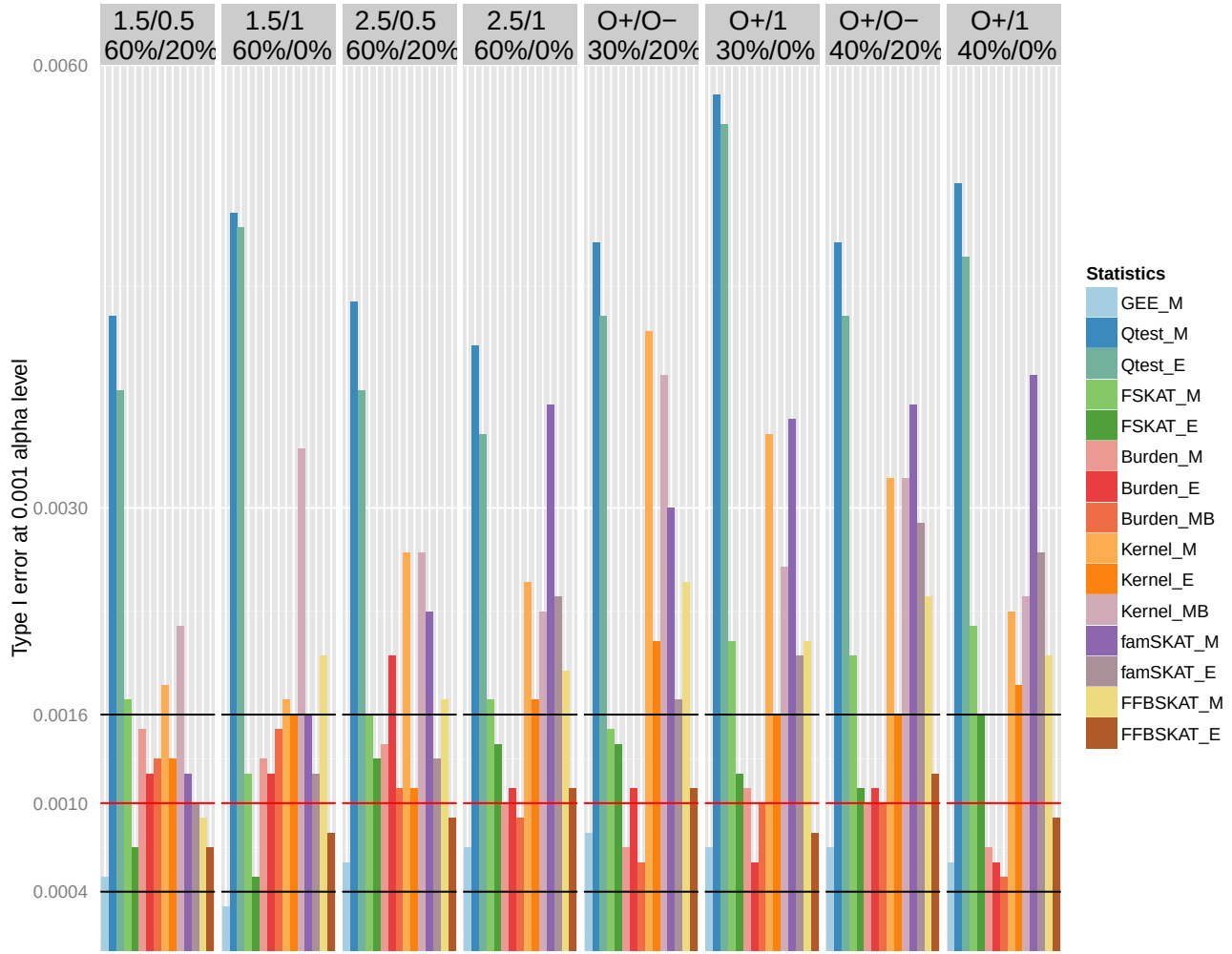


Figure 3.4: Type I error at a gene independent of the trait locus under eight trait simulation scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.001 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. The boundaries of the 95% Confidence Interval are marked out with two black lines. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights.

Figures 3.2, 3.3 and 3.4 show the Type I error for all the compared statistics at three different alpha levels. We say a statistic has inflated or deflated Type I error if its Type I error is higher or lower than the upper or lower bound of the 95% C.I., respectively; and if a statistic's Type I error is within the 95% C.I., we say it has well-behaved Type I error. First we compare the Type I error behaviors across different trait simulation scenarios. At the 0.05 alpha level, GEE_M, Qtest_M, Qtest_E, FSKAT_M, Kernel_M, and Kernel_MB have inflated Type I error. famSKAT_M and famSKAT_E, FFBSKAT_M and FFBSKAT_E have unstable behaviors, which means their Type I errors are not robust to the different trait simulation scenarios. We observed similar patterns at 0.01 and 0.001 alpha level.

Second, we compare the Type I error behaviors among different alpha levels. Compared to 0.05 alpha level, when at the 0.01 alpha level, Qtest_M and Qtest_E still have inflated Type I errors; famSKAT_M and famSKAT_E have less inflated Type I errors although they are still inflated; and FFBSKAT_M has more inflated Type I error. And when the alpha level is 0.001, Qtest_M and Qtest_E have the most inflated Type I error among all statistics. FFBSKAT_E's eight Type I errors are all within the 95% confidence interval.

Lastly, we compare the effects from assigning different weighting schemes to the markers. In general, the weights that are based on sample minor allele frequencies are tend to inflate the Type I error behavior of the statistic; while equal weights tend to help to control the Type I error. For the Burden test, the Madsen-Browning (MB) weight [Madsen and Browning, 2009] can help to control the Type I error, but for the Kernel test, it cannot help with controlling the Type I error in some scenarios.

3.4.2 Power

The naive power at the 0.05, 0.01 and 0.001 alpha levels for all compared statistics are presented in Figure 3.5 Figure 3.6 and Figure 3.7, respectively, in which the scenarios are ordered (from left to right) from one to eight. We have calculated 95% confidence intervals based on the sample size (10,000) for each statistic, which are presented as error bars. We also have labeled the statistics according to their Type I error behaviors as "d" (deflated) and "i" (inflated). The statistics that have not been labeled have well-behaved Type I error.

Recall that, in the first four scenarios, the odds ratio (OR) for risk and protective variants are fixed; while in the second four scenarios, they depend on minor allele frequencies (MAF) of the corresponding markers in the sample. The sample-MAF-dependent OR for risk variants (O^+) ranges from 2 to 10, while O^-, which is the sample-MAF-dependent OR for protective variants, ranges from 0.1 to 0.5.

In the first four scenarios (1 - 4), Figure 3.6 shows that, at the 0.01 alpha level, when the odds ratios for risk and protective markers are fixed at 1.5 and 0.5, respectively, most of the statistics have low power, especially the Burden_M and Burden_MB statistics; when the protective effects are removed while keeping the same odds ratio for risk markers, Burden_E has the highest power. And when the odds ratio for risk markers is increased to 2.5, in the presence of protective effects, Kernel_E, famSKAT_E and FFBSKAT_E have the highest power; in the absence of protective effects, Burden_E has the highest power.

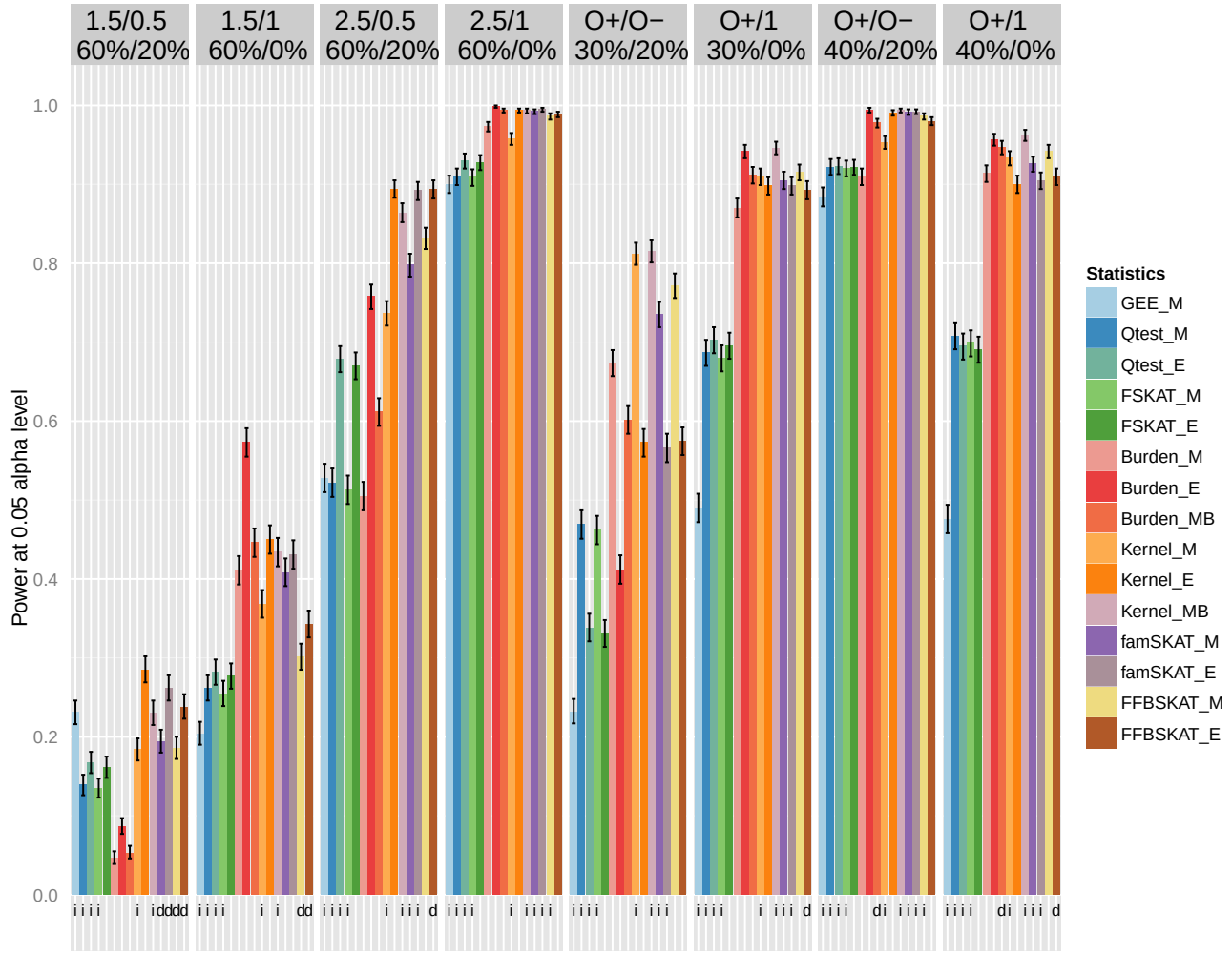


Figure 3.5: Power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.05 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights.

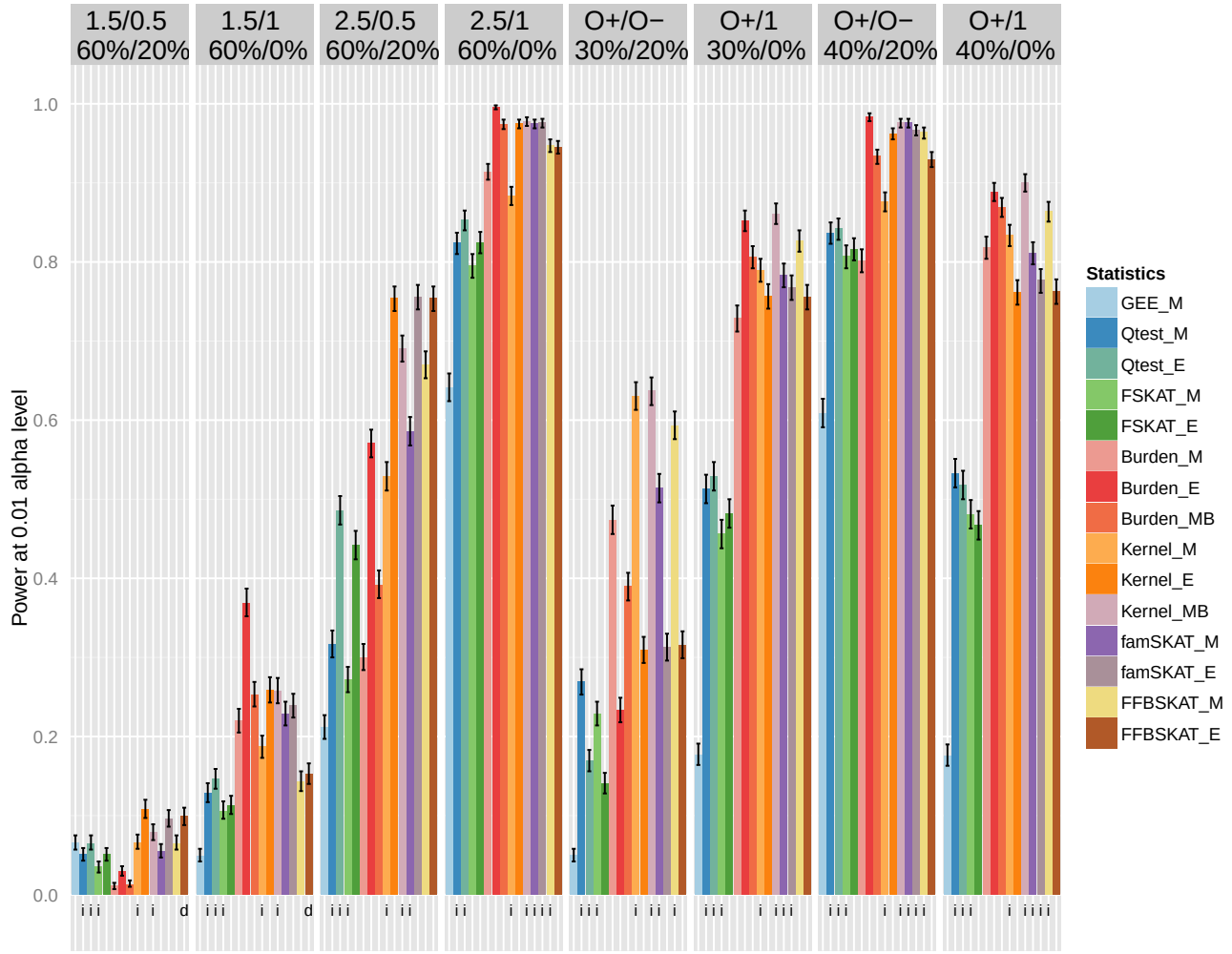


Figure 3.6: Power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.01 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights.

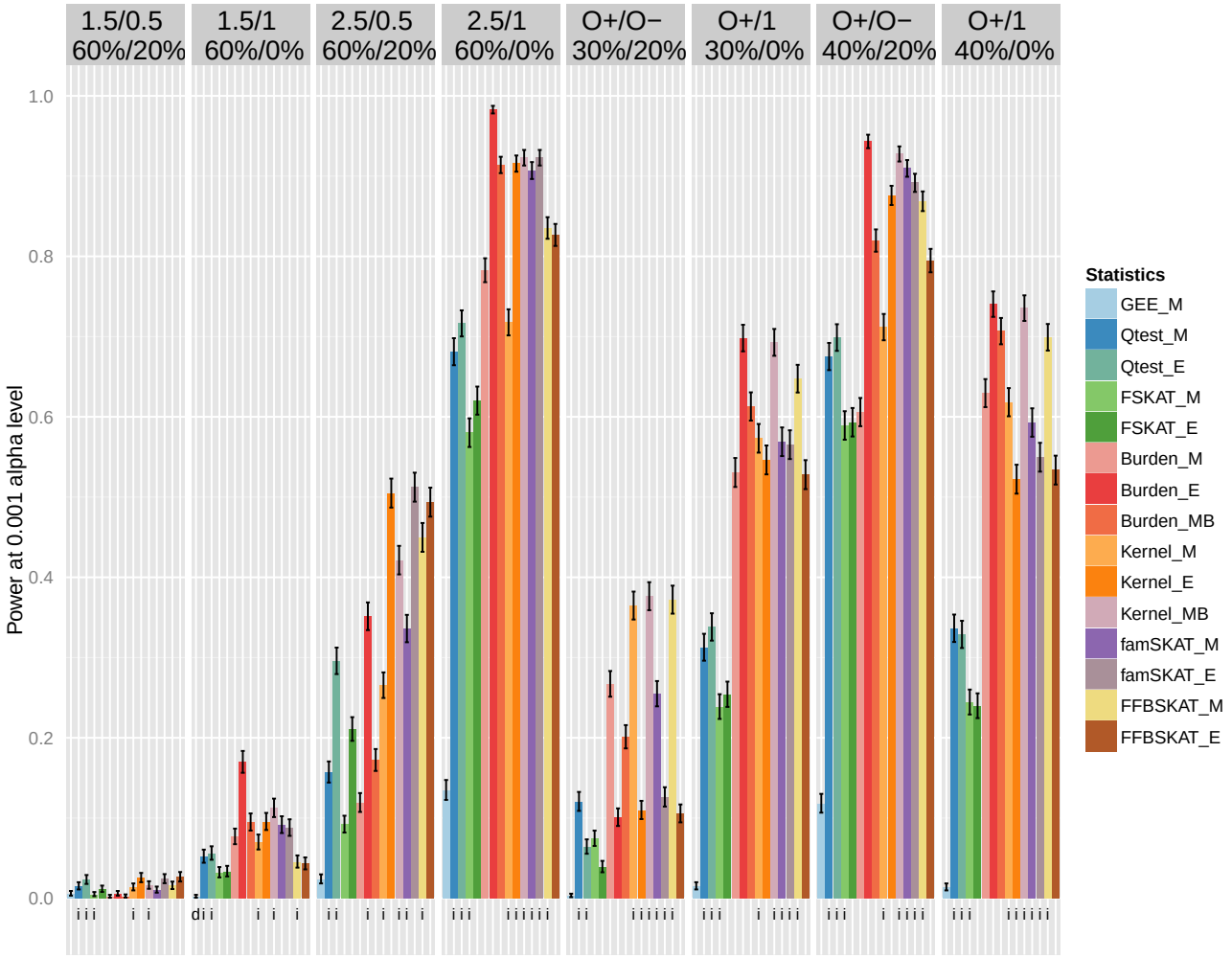


Figure 3.7: Power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.001 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights.

In the last four scenarios (5 - 8), odds ratios for both risk (O^+) and protective (O^-) variants depend on the minor allele frequencies of those rare variants: $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, and $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, in which MAF_j is the minor allele frequency for the jth marker. Figure 3.6 shows that, at 0.01 alpha level, when there are 30% risk markers and 20% protective markers, Kernel_M and Kernel_MB have the highest power; when the protective effects are removed, Burden_E and Kernel_MB have the highest power. When the percentage of risk markers is increased to 40%, in the presence of the protective effects, Burden_E has the highest power, which is different from what we have observed in Schaid et al. [2013]; in the absence of the protective effects, Kernel_MB has the highest power, but Burden_E also has very high power.

Overall, Burden_E has the highest power, and GEE_M has the lowest power. We have also observed that, statistics that assign equal weights to all the markers tend to have higher power than the statistics that assign unequal weights (e.g. sample-MAF-dependent weights) in the scenarios where the odds ratios for risk and protective markers are fixed; However, in the scenarios where the odds ratio depends on marker allele frequency, except for Burden_E in the last three scenarios, assigning sample-MAF-dependent weights or Madsen-Browning weights [Madsen and Browning, 2009] tends to improve power. Note that the power estimates are only accurate for those statistics that have good Type I error rates. The statistics that have inflated or deflated Type I error rates have overestimated or underestimated power, respectively, because, just by chance, they can produce more or fewer significant p-values than the statistics that have good Type I error rates. Recall that using unequal weights cannot better control Type I error than using equal weights, even though unequal weights match the trait simulation scenarios 5 to 8. Therefore, the high power of the statistics that use unequal weights in scenarios 5 to 8 may be due to the inflated Type I error rates. We have observed similar patterns and power behaviors for all statistics at 0.05 and 0.001 alpha levels.

We also calculated so called 'adjusted power', which means we calculated the power for each statistic by adjusting for its own Type I error rates. The results that are in Figure A2 Figure A3 and Figure A4 show that, in most scenarios, Qtest has higher power than FSKAT, and when statistics use the weights that match the underlying simulation scenarios, for ex-

ample, equal weights match with scenarios with fixed odds ratio for all the marker (scenarios 1 to 4), those statistics have higher power than the statistics that use unmatched weights, for example, using equal weights in scenarios 5 to 8, in which the odds ratios are depend on marker allele frequency.

3.5 DISCUSSION

In this chapter, we have developed a statistic, Qtest, to identify the association between rare variants and binary traits in family data by extending SKAT [Wu et al., 2011]. We have evaluated and compared the Type I error and power of this statistic together with other six ones by using family data simulated under eight different scenarios (Table 3.2). In order to simulate polygenic effects, we add the $\mathbf{P}$ in Equation 3.40, which follows a multivariate normal distribution (3.41). Note that we set σ_p at 0.38, which is calculated from the variance of $\sum_{j=1}^{m} ln(OR_j) \times g_{ij},\ i = 1 \cdots n$, the second component in Equation 3.40 where m is the total number of markers and n is sample size. We set the polygenic effect this way because in this simulation study, we focus on the effects from rare variants instead of from the polygenic part. Thus, by setting σ_p to 0.38, the simulated polygenic effects do not overwhelm the effects from rare variants. But still, setting the polygenic effect can be arbitrary, and it may depend on the purpose of the simulation study, for example in Yan et al. [2015].

Table 3.2 shows the average number of variants by minor allele frequency (MAF). However, since each replicate contains different samples of haplotypes from the haplotype pool, not all the risk or protective variants are polymorphic in each of the sampled datasets. Thus, sometimes, non-polymorphic risk or protective variants have been dropped from the sampled datasets during the simulation. In other words, the average number of variants by MAF just gives a general distribution of those rare variants within the selected region. It is unknown to the researcher that whether all or part of them are risk and/or protective, just like one would expect in a real data analysis. In this study, we have included two statistics that are designed for quantitative traits only, famSKAT and FFBSKAT. However, we are applying them on binary traits to study their performances when treating a binary trait as a "0,

1" continuous variable with "1" represents an affected individual. Although it is kind of counter-intuitive, FFBSKAT showed well-behaved Type I error and power in most of the scenarios, while famSKAT has inflated Type I errors in most of the scenarios at 0.05 alpha level.

3.5.1 Weighting matrix

When applying association test statistics on rare variants, one can assign different weights for the markers according to their MAF or some prior knowledge. In our simulation, we have used equal weights, sample-MAF-dependent weights for all statistics except for GEE, which has issues setting equal weights in its R implementation, and Madsen-Browning weights [Madsen and Browning, 2009] for the Kernel and the Burden test statistics. We have compared the effects of these different weighting schemes, which do influence the Type I error behaviors and power of the statistics (e.g. Figure 3.3, Figure 3.6). When analyzing real data, using equal weights has some advantages: first, assume we do not know whether or not a marker with small MAF, thus, large effect size, is risk or protective or neutral; Second, it is inappropriate to assign weights based on sample MAF (MAF calculated based on sampled families), because the odds ratio or effect sizes have been simulated based on population MAF (MAF calculated based on haplotype pool). Third, in a family-based dataset, the sample MAFs are quite different from the MAFs in a population-based dataset, thus quite different from population MAFs, and last but not the least, population MAF has been distorted by the ascertainment procedure in sampled dataset. Therefore, using family-based sample MAF to calculate weights could introduce bias, and by using equal weights, we are trying to control for all these potential bias. However, whether or not using equal weights on real data is a good idea depends on the unknown true state of nature. So, the decision of which weighting scheme to use has to be driven by the prior beliefs, which means if one has prior knowledge of the markers, one can use the population-MAF-dependent weights for the markers, which may improve the power. Or, one could use both equal and MAF-dependent weights to analyze the data at the expense of multiple testing.

3.5.2 Inflated Type I error

Qtest and FSKAT are very similar to each other, and they have similar power behaviors across different scenarios, but they have different Type I error behaviors. Especially, Qtest has inflated Type I error at 0.001 alpha level (Figure 3.4). This may be due to the different approximation methods used in the different statistics. Specifically, recall that the variance component τ is tested at the boundary, which makes the test statistic not follow a standard chi-square distribution with one degree of freedom. Rather, it follows a mixture of chi-square distributions. FSKAT can calculate the weights for the mixture of chi-square distributions (λ in Equation (3.42)), while Qtest applies the Laplace approximation method, and uses a scaled chi-square distribution (Equation (3.39)) to approximate the mixture of chi-square distributions. Therefore, compared to FSKAT, Qtest has an extra layer of approximation, which might be one of the reasons why Qtest does not have well calibrated Type I errors especially in the tail (alpha = 0.001). For Qtest, more theoretical derivations are needed to remove this layer so as to instead use the mixture of chi-square distributions as the null distribution of the statistic.

Another possible explanation of the inflation of Type I errors for Qtest is that, as mentioned in Lin [1997], when the data are binary, the sample size and the number of levels of each random effect is small, the performance of Laplace approximation is unsatisfactory. When the sample size and the number of levels of each random effect increase, the accuracy of Laplace-based approximation methods could quickly improve, thus improve the score test. Therefore, a Monte Carlo based simulation method such as importance sampling may be needed when dealing with binary traits to improve performance. This could be part of the future work.

Moreover, the inflated Type I error behaviors, not only for Qtest, but also for other statistics, for example famSKAT, may be also caused by ascertainment procedure. In this study, we have ascertained the simulated families according to the criteria that the family should contain at least two affected individuals in the youngest generation. The ascertainment procedure has resulted in increased portion of affected subjects in the sample from the prevalence (5%) under null hypothesis, and introduces selection bias [Clark et al., 2005; Siegmund and

Langholz, 2002]. Except for the Burden and the Kernel tests, other statistics assume the families are randomly ascertained. Thus, treating ascertained sample as completely random without any adjustment is not quite appropriate. Without proper adjustment, the estimated variance components are biased, influencing test statistics [Oualkacha et al., 2013] by inflating the Type I error. Therefore, it is necessary to adjust for the ascertainment procedure somewhere in the statistic if the families are not randomly ascertained to obtain a better control of Type I error behavior, which could be very challenging [Vieland and Hodge, 1995].

For the Burden and the Kernel tests, with different weighting schemes (e.g. Equal weights, sample-MAF-dependent weights), we have observed that they have well-behaved Type I error as well as high power across most of the simulated scenarios, especially for the Burden tests. Note that these statistics are constructed in a retrospective way such that they assume the traits are fixed, instead of random. This way, these statistics could avoid modeling the ascertainment procedure and obtain well-behaved Type I error behaviors. Note that the inflated Type I error behaviors for famSKAT and FFBSKAT could also be due to treating binary traits as continuous. For GEE, in Wang et al. [2013], GEE has showed well-behaved Type I error in both random and ascertainment sampling designs, which increases its robustness to ascertainment bias. But GEE has low power, which could be a trade off of this robustness. In real studies, the multi-generation family data could be collected retrospectively with some ascertainment criteria, or collected prospectively as the study goes on such as the family data in Framingham heart study. Obviously, collecting data prospectively could cost a lot more time and money than the retrospective way. One possible way to avoid the ascertainment issue is ascertaining families using the secondary traits while analyzing the primary traits, but it only works when these two are not highly associated [Schifano et al., 2012], as also shown in De Andrade and Amos [2000], the ascertainment bias had ignorable effect when the correlation between primary and secondary traits are ignorable. Also, the statistics may have low power when analyzing the primary traits.

3.5.3 Population stratification and family structure mis-specification

In real data analysis, researchers have to deal with confounding factors such as population stratification, which results in different allele frequencies in sampled population due to a variety of reasons such as different ancestors. The statistics in our simulation are not robust to population stratification by themselves. However, one can apply the Principle Component Analysis (PCA) method on founders of the pedigrees and check for clusters, or to adjust for population stratification [Zhu et al., 2008; Liu et al., 2013b] and add into the model as a covariate. In this simulation study, we do not simulate population stratified data, which would be carried out in future work.

However, before checking for population stratification using PCA method, one has to make sure the family structures are correctly specified. GEE [Wang et al., 2013], although has lower power than any other statistics in our study in most scenarios, in theory, is robust to mis-specification of family structure as it uses working correlation structure and has valid estimates of mean and variance. But we need more simulation with mis-specified family structure to check that. Also, by genotyping a large number of individuals, one can identify the relationship within families and thus check for mis-specified family structure.

3.5.4 Untyped subjects

Another issue in real data analysis is missing phenotype or genotype data. For missing phenotype data, one can first calculate kinship matrix, and then drop the subjects with missing phenotypes from the dataset and also from the kinship matrix. For missing genotype data, although in this simulation, we do not address this issue, there are many imputation methods that can be applied. For example, BEAGLE [Browning, 2006], MACH [Li et al., 2006], IMPUTE [Marchini et al., 2007], GIGI [Cheung et al., 2013], which can impute missing genotypes in family-based data. Some statistical algorithms for imputing genotypes within families, which are based on Lander Green [Lander and Green, 1987] or Elston-Stewart [Elston and Stewart, 1971] algorithms, or Monte Carlo sampling [Heath, 1997; Lange and Sobel, 1996], are described in Chen and Abecasis [2007] and Visscher and Duffy [2006], and implemented in MERLIN [Abecasis et al., 2002; Abecasis and Wigginton, 2005], MENDEL

[Lange et al., 2001] and some other programs. In family-based data, missing genotype imputation has to condition on Mendelian consistency [Cheung et al., 2014], population stratification and the correlation among markers, which can be very complicated for rare variants. In future work, we would like to focus on the development of a missing genotype imputation method on family-based data and integrate it into Qtest.

4.0 SUMMARY AND FUTURE WORK

4.1 A SIMULATION-BASED COMPARATIVE STUDY OF FAMILY-BASED ASSOCIATION TESTS

In Chapter 2, we compared many single-common-variant association analysis statistics, and discussed the advantages and disadvantages when applying these statistics to real data. During the preparation of this dissertation, many new statistics have been developed (e.g. ROADTRIPS [Thornton and McPeek, 2010]). Thus, we would like to continue comparing single-common-variant association statistics by including those newly developed ones. Also, in our study, we did not evaluate or compare the statistics' performance in the presence of population stratification. Thus, we would like to modify our simulation to mimic the population stratification in simulated family-based data. To do this, we could first simulate a set of pedigrees with one set of allele frequencies, and simulate another set of pedigrees with a different set of allele frequencies. Then we could combine these two sets of family data to form one dataset, in which there is population stratification reflected by the difference of underlying allele frequencies. Moreover, we would like to compare all these statistics on real data to further evaluate their performances.

4.2 FAMILY-BASED RARE VARIANTS ASSOCIATION ANALYSIS FOR BINARY TRAITS

In Chapter 3, we developed a statistic, Qtest, to test for association between a binary trait and rare variants on family-based data by extending SKAT [Wu et al., 2011]. Through

simulation, we also found that this method has elevated Type I error behavior (Figures 3.2 , 3.3 and 3.4), and not as powerful as the statistics (Kernel and Burden) developed by Schaid et al. [2013] (Figures 3.7, 3.6 and 3.5).

Therefore, for future work, we would like to improve this statistic in the following perspectives:

1) According to Lin [1997], the elevated Type I error may be caused by non-accurate approximation from the Laplace method due to insufficient sample size. Lin [1997] also suggested combining the Laplace approximation method with Monte Carlo importance sampling or similar method to improve the approximation accuracy. Booth and Hobert [1999] and McCulloch [1997] have provided methods of applying Monte Carlo Expectation-Maximization (MCEM) on Generalized Linear Mixed Models (GLMM). In general, MCEM uses Monte Carlo procedure to enlarge the sample size in simulation, and treats the random effects in the GLMM as missing data, and applies EM methods to obtain estimates of interested parameters. Thus, we would like to try integrate this MCEM method into Qtest to improve its performance. Papachristou et al. [2011] have applied this algorithm to an association study for common variants, and they have found that this method needs a burn-in step, which although very time consuming, would not be a problem as the speed of computers is constantly improving.

2) Qtest has no adjustment for ascertainment, and it assumes the sampled families are randomly selected from the population. Thus, ascertainment bias [Clark et al., 2005; Siegmund and Langholz, 2002] affects the behaviors of Qtest and other similar statistics that based on mixed model [Oualkacha et al., 2013]. There are some methods to adjust for that bias, for example, the methods proposed in Schaid et al. [2013] assume that the trait is fixed, instead of a random variable, thus do not assume the sampled families are randomly selected from the population. We have seen that in Chapter 3, these two methods (Burden and Kernel) have better Type I error and power than Qtest.

For Qtest, one possible way to adjust for ascertainment bias is to construct a likelihood that is condtioned on ascertainment. Recall in Chapter 2, the Likelihood ratio test based on Generalized Linear Penetrance Model [Lange et al., 2005] implemented in Mendel [Lange et al., 2001], PENE, has been evaluated. This method constructs a log-likelihood for

the ascertainment procedure and subtracts it from the log-likelihood for the pedigree. In other words, if we define the joint likelihood of the pedigree and ascertainment procedure as L(Pedigree, Ascertainment), then we might be able to construct a likelihood for the ascertainment procedure L(Ascertainment), which could be a function of or proportional to the multiplication of the probability of observing, for example, two affected offspring in each family. Then, we might be able to construct the conditional likelihood L(Pedigree | Ascertainment) as L(Pedigree, Ascertainment)/L(Ascertainment). After taking logarithm, it becomes: log(L(Pedigree | Ascertainment)) = log(L(Pedigree, Ascertainment)) - log(Ascertainment). Finally, we might be able to use the conditional log-likelihood, log(L(Pedigree | Ascertainment)), as our likelihood and derive the score statistic and information matrix. Another way to correct for ascertainment bias is to ascertain according to one trait that is not of interested, then analyze the trait of interested, given that the two traits are not highly correlated. This is more of a study design issue, and it may cause power loss when analyzing the trait of interest. We would like to evaluate this study design in the future, too.

3) In this study, we have not simulated rare variant family-based data in the presence of population stratification. We would like to evaluate the performance of the statistics we have compared in Chapter 3 in the presence of population stratification. In order to adjust for population stratification in Qtest, we could apply the strategy proposed by Zhu et al. [2008]; Liu et al. [2013b], in which the population stratification has been identified and added into the model as a covariate. However, this method may not work well on rare variants. Or, we could first detect population stratification by using the PCA method or the method proposed in Qiao et al. [2013] on pedigree founders, and separately analyze the data. Obviously, this method would reduce sample size.

4) In Chapter 3, we do not simulate any family-based data for X-chromosome rare variants, which requires recoded genotypes for males and re-calculated kinship coefficients because males have only one copy of X-chromosome. There are two recoding methods that can be applied to recode genotypes for males. One is developed by Zheng et al. [2007] and the other one is developed by Clayton [2008]. And these two recoding methods have been compared for common variants by Loley et al. [2011] and Konig et al. [2014]. For future work, we will apply both methods and compare them for rare variant analysis.

5) In our simulation, we do not simulate any scenarios containing untyped individual. In other words, Qtest assumes the data are complete, every individual has fully typed genotype. In real data analysis, one has to handle untyped individuals. One simple way is to drop them, but that would cause reduced sample size. So, it would be better if we can fill in the missing genotype by constructing a proper imputation procedure in family-based data for rare variants on both autosomes and X-chromosome. For missing phenotype, we suggest to first calculate the kinship matrix based on the complete family structure, and then drop the subject who has missing phenotype from the dataset and the kinship matrix.

In family-based data, the genotypes within each family should be Mendelianly consistent when assuming no mutation, which is appropriate in small samples. One popular imputation method in population-based data, which assumes independence among individuals, is to fill in the missing genotype by sampling from the pool of candidate genotypes in the population according to their frequencies. However, in family-based imputation, due to Mendelian consistency, the number of candidate genotypes in the pool are limited. Thus, the first step of imputing the missing genotype in family-based data is to identify correct candidate genotypes by checking for Mendelian inconsistencies within a family. Then, one can sample genotype from the limited or Mendelianly consistent candidate genotypes and fill in the missing genotype. This shall be done individual by individual and family by family. Then, one can apply Qtest or other statistics on imputed dataset to test for association.

6) There is a newly developed method by Zhang et al. [2014], Weighted Sum Mixed Model (WSMM), which applies a permutation methods on family data to obtain adjusted weights for association analysis between quantitative traits and rare variants. It was compared to famSKAT, and the author claimed that it also can be applied on binary traits. So we would like to include this statistic into our study in the future.

APPENDIX A

SUPPLEMENTARY TABLES AND FIGURES

Table A1: Type I error and power for all statistics across all scenarios.

Statistics	Family Structure	No Linkage			Complete Linkage		
		Dom	Add	Rec	Dom	Add	Rec
Test for association (Null A)							
ALLELE_FREQ	2gen	0.050	0.046	0.044	0.050	0.046	0.045
	3gen	0.047	0.057	0.063	0.043	0.056	0.063
	2genUP	0.045	0.035	0.060	0.047	0.040	0.061
	3genUG	0.055	0.060	0.065	0.050	0.060	0.063
	3genUGP	0.054	0.059	0.054	0.052	0.048	0.052
CACO_FISHER	2gen	0.037	0.044	0.049	0.040	0.045	0.044
	3gen	0.058	0.055	0.045	0.075	0.062	0.056
	2genUP	0.059	0.040	0.039	0.033	0.044	0.048
	3genUG	0.054	0.055	0.052	0.069	0.057	0.047
	3genUGP	0.048	0.043	0.043	0.074	0.049	0.045
CACO_ZMAX	2gen	0.041	0.046	0.047	0.038	0.038	0.045
	3gen	0.054	0.042	0.046	0.072	0.052	0.056
	2genUP	0.048	0.038	0.043	0.034	0.047	0.063
	3genUG	0.041	0.048	0.041	0.062	0.043	0.048

Table A1 Continued							
	Family	No Linkage			Complete Linkage		
Statistics	Structure	Dom	Add	Rec	Dom	Add	Rec
	3genUGP	0.053	0.038	0.040	0.077	0.038	0.035
PENE	2gen	0.025	0.019	0.028	0.029	0.021	0.034
	3gen	0.064	0.045	0.035	0.089	0.064	0.040
	2genUP	0.030	0.028	0.033	0.026	0.028	0.038
	3genUG	0.062	0.044	0.043	0.078	0.055	0.049
	3genUGP	0.052	0.044	0.042	0.080	0.065	0.051
LME	2gen	0.018	0.041	0.054	0.022	0.051	0.046
	3gen	0.107	0.126	0.128	0.145	0.140	0.137
	2genUP	0.042	0.037	0.046	0.032	0.045	0.061
	3genUG	0.022	0.011	0.028	0.027	0.016	0.024
	3genUGP	0.036	0.044	0.033	0.063	0.056	0.047
GEE_ind	2gen	0.042	0.043	0.064	0.048	0.053	0.059
	3gen	0.079	0.067	0.057	0.082	0.083	0.065
	2genUP	0.051	0.047	0.060	0.045	0.055	0.055
	3genUG	0.088	0.062	0.070	0.090	0.070	0.078
	3genUGP	0.081	0.069	0.072	0.081	0.078	0.068
GEE_ex	2gen	0.055	0.064	0.079	0.060	0.065	0.086
	3gen	0.085	0.074	0.078	0.092	0.096	0.071
	2genUP	0.069	0.086	0.094	0.071	0.088	0.103
	3genUG	0.093	0.080	0.083	0.082	0.088	0.089
	3genUGP	0.093	0.087	0.089	0.091	0.091	0.082
g_gee1	2gen	0.043	0.044	0.067	0.047	0.053	0.062
	3gen	0.073	0.071	0.057	0.080	0.081	0.069
	2genUP	0.049	0.045	0.059	0.045	0.053	0.057
	3genUG	0.088	0.061	0.066	0.084	0.067	0.073
	3genUGP	0.079	0.067	0.069	0.077	0.076	0.064

Table A1 Continued							
	Family	No Linkage			Complete Linkage		
Statistics	Structure	Dom	Add	Rec	Dom	Add	Rec
Transmit	2gen	0.054	0.057	0.047	0.058	0.070	0.091
	3gen	0.052	0.052	0.055	0.080	0.065	0.063
	2genUP	0.057	0.047	0.058	0.058	0.076	0.092
	3genUG	0.043	0.052	0.039	0.075	0.073	0.054
	3genUGP	0.045	0.050	0.049	0.082	0.073	0.068
Transmit_r	2gen	0.059	0.058	0.045	0.057	0.047	0.053
	3gen	0.052	0.057	0.062	0.066	0.060	0.059
	2genUP	0.059	0.049	0.060	0.046	0.057	0.064
	3genUG	0.043	0.065	0.038	0.066	0.061	0.053
	3genUGP	0.053	0.052	0.048	0.078	0.077	0.057
QTDT_ad	2gen	0.034	0.047	0.048	0.041	0.039	0.040
	3gen	0.032	0.035	0.029	0.028	0.031	0.031
	2genUP	0.043	0.039	0.042	0.032	0.037	0.048
	3genUG	0.021	0.025	0.027	0.025	0.036	0.036
	3genUGP	0.015	0.032	0.029	0.021	0.031	0.022
MM1	2gen	0.288	0.301	0.310	0.268	0.295	0.298
	3gen	0.374	0.404	0.416	0.430	0.473	0.434
	2genUP	0.276	0.328	0.271	0.287	0.295	0.282
	3genUG	0.383	0.411	0.436	0.395	0.415	0.452
	3genUGP	0.360	0.383	0.382	0.361	0.392	0.407
Test for association in the absence of linkage (Null AL)							
PMDom_LD|NL	2gen	0.067	0.062	0.075	0.060	0.063	0.075
	3gen	0.056	0.063	0.049	0.052	0.063	0.052
	2genUP	0.076	0.061	0.073	0.064	0.050	0.058
	3genUG	0.056	0.058	0.059	0.058	0.061	0.061
	3genUGP	0.042	0.052	0.031	0.042	0.043	0.030

Table A1 Continued							
	Family	No Linkage			Complete Linkage		
Statistics	Structure	Dom	Add	Rec	Dom	Add	Rec
PMRec_LD|NL	2gen	0.058	0.069	0.067	0.048	0.067	0.065
	3gen	0.055	0.067	0.051	0.056	0.064	0.052
	2genUP	0.063	0.075	0.068	0.048	0.054	0.034
	3genUG	0.073	0.060	0.062	0.071	0.057	0.057
	3genUGP	0.060	0.061	0.063	0.062	0.064	0.070
PMMbase_LD|NL	2gen	0.056	0.055	0.066	0.052	0.051	0.067
	3gen	0.069	0.053	0.056	0.067	0.057	0.051
	2genUP	0.058	0.042	0.064	0.045	0.047	0.064
	3genUG	0.067	0.057	0.067	0.078	0.065	0.064
	3genUGP	0.067	0.055	0.042	0.055	0.053	0.046
Test for association in the presence of linkage (Null CL)							
FBAT_e	2gen	0.052	0.053	0.042	0.052	0.044	0.051
	3gen	0.040	0.051	0.052	0.047	0.038	0.043
	2genUP	0.051	0.041	0.039	0.042	0.047	0.052
	3genUG	0.037	0.044	0.036	0.044	0.052	0.053
	3genUGP	0.032	0.037	0.038	0.044	0.046	0.040
AS|LINK	2gen	0.019	0.021	0.032	0.028	0.027	0.037
	3gen	0.033	0.030	0.022	0.047	0.034	0.024
	2genUP	0.008	0.017	0.019	0.009	0.029	0.019
	3genUG	0.033	0.025	0.031	0.044	0.027	0.038
	3genUGP	0.026	0.022	0.014	0.032	0.020	0.019
PMDom_LD|L	2gen	0.077	0.087	0.101	0.062	0.071	0.084
	3gen	0.062	0.076	0.072	0.073	0.079	0.073
	2genUP	0.110	0.093	0.111	0.076	0.078	0.077
	3genUG	0.072	0.081	0.086	0.078	0.088	0.084
	3genUGP	0.070	0.079	0.058	0.068	0.066	0.047

Table A1 Continued							
	Family	No Linkage			Complete Linkage		
Statistics	Structure	Dom	Add	Rec	Dom	Add	Rec
PMRec_LD|L	2gen	0.069	0.085	0.094	0.063	0.074	0.053
	3gen	0.065	0.073	0.058	0.064	0.072	0.059
	2genUP	0.086	0.101	0.101	0.060	0.072	0.055
	3genUG	0.090	0.085	0.079	0.087	0.068	0.084
	3genUGP	0.073	0.079	0.090	0.079	0.077	0.095
PMMbase_LD|L	2gen	0.062	0.062	0.091	0.062	0.067	0.081
	3gen	0.074	0.070	0.073	0.074	0.070	0.072
	2genUP	0.069	0.067	0.078	0.061	0.068	0.092
	3genUG	0.078	0.071	0.090	0.087	0.079	0.083
	3genUGP	0.084	0.066	0.056	0.068	0.075	0.071
Test for association or linkage (Null NL)							
QTDT_ar	2gen	0.049	0.051	0.046	0.041	0.041	0.045
	3gen	0.046	0.041	0.053	0.069	0.069	0.060
	2genUP	NaN	NaN	NaN	NaN	NaN	NaN
	3genUG	0.040	0.043	0.050	0.060	0.061	0.058
	3genUGP	0.032	0.035	0.042	0.072	0.040	0.053
FBAT	2gen	0.054	0.057	0.047	0.058	0.070	0.091
	3gen	0.052	0.052	0.055	0.080	0.065	0.063
	2genUP	0.053	0.039	0.049	0.045	0.063	0.063
	3genUG	0.043	0.054	0.040	0.071	0.071	0.056
	3genUGP	0.044	0.043	0.051	0.067	0.064	0.064
GC1	2gen	0.055	0.057	0.047	0.060	0.070	0.091
	3gen	0.052	0.053	0.057	0.081	0.067	0.065
	2genUP	0.055	0.046	0.056	0.053	0.070	0.087
	3genUG	0.044	0.054	0.042	0.077	0.076	0.056
	3genUGP	0.047	0.043	0.041	0.084	0.071	0.065

Table A1 Continued							
	Family	No Linkage			Complete Linkage		
Statistics	Structure	Dom	Add	Rec	Dom	Add	Rec
GC2	2gen	0.055	0.057	0.047	0.060	0.070	0.091
	3gen	0.052	0.053	0.057	0.081	0.067	0.065
	2genUP	0.058	0.047	0.059	0.058	0.076	0.091
	3genUG	0.044	0.054	0.042	0.077	0.076	0.056
	3genUGP	0.050	0.043	0.041	0.086	0.075	0.066
GC1CT	2gen	0.051	0.049	0.045	0.042	0.046	0.051
	3gen	0.050	0.045	0.051	0.068	0.080	0.056
	2genUP	0.048	0.055	0.043	0.045	0.044	0.041
	3genUG	0.036	0.045	0.049	0.074	0.050	0.055
	3genUGP	0.047	0.052	0.046	0.081	0.063	0.058
GC2CT	2gen	0.051	0.049	0.045	0.042	0.046	0.051
	3gen	0.050	0.045	0.051	0.068	0.080	0.056
	2genUP	0.048	0.055	0.043	0.045	0.044	0.041
	3genUG	0.036	0.045	0.050	0.076	0.052	0.056
	3genUGP	0.047	0.053	0.047	0.083	0.063	0.059
Mendel_TDT	2gen	0.045	0.041	0.035	0.047	0.057	0.076
	3gen	0.039	0.043	0.040	0.068	0.044	0.051
	2genUP	NaN	NaN	NaN	NaN	NaN	NaN
	3genUG	0.030	0.038	0.032	0.058	0.046	0.046
	3genUGP	0.027	0.015	0.017	0.039	0.027	0.028
g_tdt	2gen	0.054	0.057	0.047	0.060	0.070	0.091
	3gen	0.052	0.052	0.055	0.080	0.065	0.063
	2genUP	NaN	NaN	NaN	NaN	NaN	NaN
	3genUG	0.038	0.050	0.046	0.072	0.067	0.062
	3genUGP	0.043	0.039	0.040	0.063	0.049	0.051
g_1tdt	2gen	0.054	0.053	0.042	0.052	0.044	0.051
	3gen	0.040	0.051	0.052	0.047	0.038	0.045

Table A1 Continued							
	Family	No Linkage			Complete Linkage		
Statistics	Structure	Dom	Add	Rec	Dom	Add	Rec
	2genUP	0.048	0.039	0.059	0.058	0.043	0.052
	3genUG	0.033	0.054	0.039	0.042	0.049	0.052
	3genUGP	0.046	0.049	0.044	0.053	0.047	0.048
g_pdt	2gen	0.039	0.045	0.055	0.039	0.053	0.038
	3gen	0.037	0.035	0.054	0.040	0.042	0.051
	2genUP	0.048	0.045	0.044	0.043	0.049	0.056
	3genUG	0.028	0.054	0.036	0.037	0.048	0.044
	3genUGP	0.038	0.040	0.034	0.050	0.052	0.046
GDT	2gen	0.036	0.052	0.060	0.037	0.048	0.048
	3gen	0.050	0.045	0.049	0.085	0.067	0.051
	2genUP	0.050	0.043	0.052	0.039	0.050	0.069
	3genUG	0.043	0.047	0.054	0.075	0.060	0.063
	3genUGP	0.046	0.039	0.047	0.077	0.068	0.047
poGDT	2gen	0.047	0.049	0.053	0.056	0.040	0.063
	3gen	0.046	0.050	0.050	0.077	0.068	0.062
	2genUP	0.046	0.039	0.055	0.040	0.050	0.069
	3genUG	0.041	0.046	0.038	0.045	0.063	0.056
	3genUGP	0.038	0.053	0.046	0.052	0.054	0.055
MQLStest_caco	2gen	0.035	0.037	0.048	0.041	0.042	0.048
	3gen	0.056	0.042	0.041	0.075	0.064	0.045
	2genUP	0.044	0.043	0.055	0.043	0.049	0.053
	3genUG	0.050	0.045	0.049	0.066	0.053	0.059
	3genUGP	0.042	0.046	0.041	0.062	0.061	0.052
WQLS_r	2gen	0.048	0.050	0.056	0.052	0.065	0.077
	3gen	0.048	0.052	0.058	0.081	0.069	0.060
	2genUP	0.047	0.041	0.052	0.037	0.046	0.067
	3genUG	0.049	0.048	0.047	0.066	0.066	0.062

Table A1 Continued							
	Family	No Linkage			Complete Linkage		
Statistics	Structure	Dom	Add	Rec	Dom	Add	Rec
	3genUGP	0.043	0.053	0.044	0.077	0.070	0.063
MQLStest_r	2gen	0.036	0.050	0.049	0.050	0.055	0.070
	3gen	0.064	0.049	0.049	0.073	0.063	0.046
	2genUP	0.050	0.038	0.058	0.043	0.058	0.050
	3genUG	0.055	0.049	0.055	0.064	0.051	0.058
	3genUGP	0.048	0.056	0.040	0.064	0.056	0.052
MQLS_e	2gen	0.036	0.054	0.053	0.046	0.055	0.064
	3gen	0.066	0.052	0.049	0.079	0.062	0.049
	2genUP	0.053	0.038	0.059	0.039	0.056	0.048
	3genUG	0.055	0.046	0.057	0.065	0.056	0.058
	3genUGP	0.047	0.055	0.043	0.066	0.055	0.053
IQLS	2gen	0.036	0.054	0.053	0.046	0.055	0.064
	3gen	0.066	0.052	0.049	0.079	0.062	0.049
	2genUP	0.053	0.038	0.059	0.039	0.056	0.048
	3genUG	0.055	0.046	0.057	0.065	0.056	0.058
	3genUGP	0.047	0.055	0.043	0.066	0.055	0.053
g_mqls	2gen	0.036	0.054	0.054	0.048	0.056	0.064
	3gen	0.066	0.052	0.050	0.079	0.062	0.050
	2genUP	0.053	0.038	0.059	0.040	0.056	0.051
	3genUG	0.057	0.047	0.057	0.065	0.056	0.059
	3genUGP	0.048	0.055	0.045	0.067	0.055	0.054
g_qlsw	2gen	0.046	0.052	0.052	0.059	0.066	0.088
	3gen	0.067	0.060	0.051	0.077	0.067	0.048
	2genUP	0.053	0.042	0.057	0.039	0.051	0.070
	3genUG	0.066	0.050	0.048	0.069	0.066	0.062
	3genUGP	0.040	0.054	0.048	0.072	0.054	0.051

Table A1 Continued							
	Family	No Linkage			Complete Linkage		
Statistics	Structure	Dom	Add	Rec	Dom	Add	Rec
PMDom_LDL	2gen	0.063	0.075	0.079	0.158	0.156	0.225
	3gen	0.066	0.072	0.061	0.197	0.124	0.093
	2genUP	0.082	0.072	0.079	0.117	0.140	0.207
	3genUG	0.069	0.068	0.069	0.150	0.112	0.108
	3genUGP	0.051	0.071	0.057	0.140	0.116	0.075
PMRec_LDL	2gen	0.064	0.082	0.067	0.120	0.187	0.354
	3gen	0.067	0.072	0.056	0.127	0.127	0.097
	2genUP	0.067	0.086	0.086	0.096	0.148	0.219
	3genUG	0.086	0.089	0.070	0.132	0.117	0.119
	3genUGP	0.058	0.066	0.072	0.116	0.111	0.112
PMMbase_LDL	2gen	0.053	0.061	0.069	0.050	0.051	0.062
	3gen	0.077	0.052	0.062	0.221	0.111	0.084
	2genUP	0.059	0.065	0.064	0.057	0.061	0.064
	3genUG	0.064	0.058	0.070	0.191	0.094	0.097
	3genUGP	0.067	0.069	0.045	0.132	0.103	0.079

Note: Due to different null hypotheses, blue colored values are power, others are Type I error. Family structures: 2gen: fully typed two-generation families; 3gen: fully typed three-generation families; 2genUP: two-generation families with one untyped parent; 3genUG: three-generation families with two untyped grandparents; 3genUGP: three-generation families with two untyped grandparents and some untyped parents.

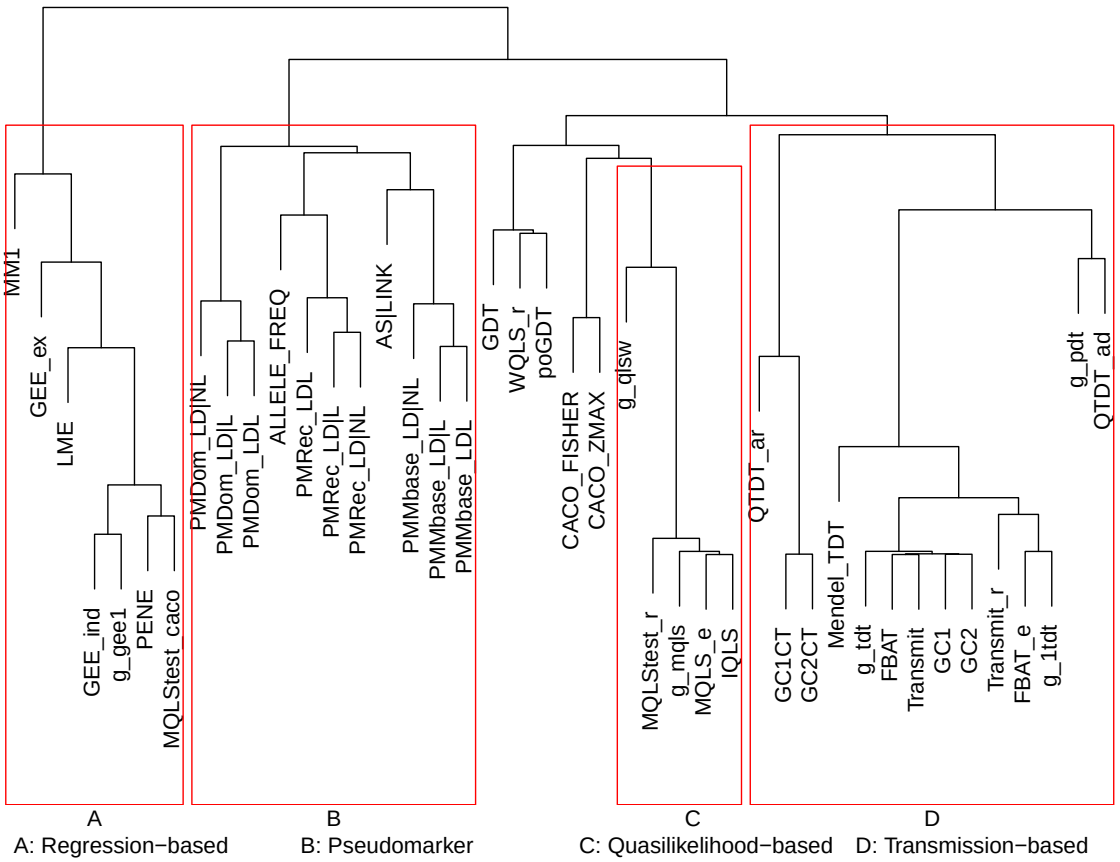


Figure A1: Hierarchical clustering plot based on Euclidean distance of p-values under Null NL across fully typed family structures and penetrance models

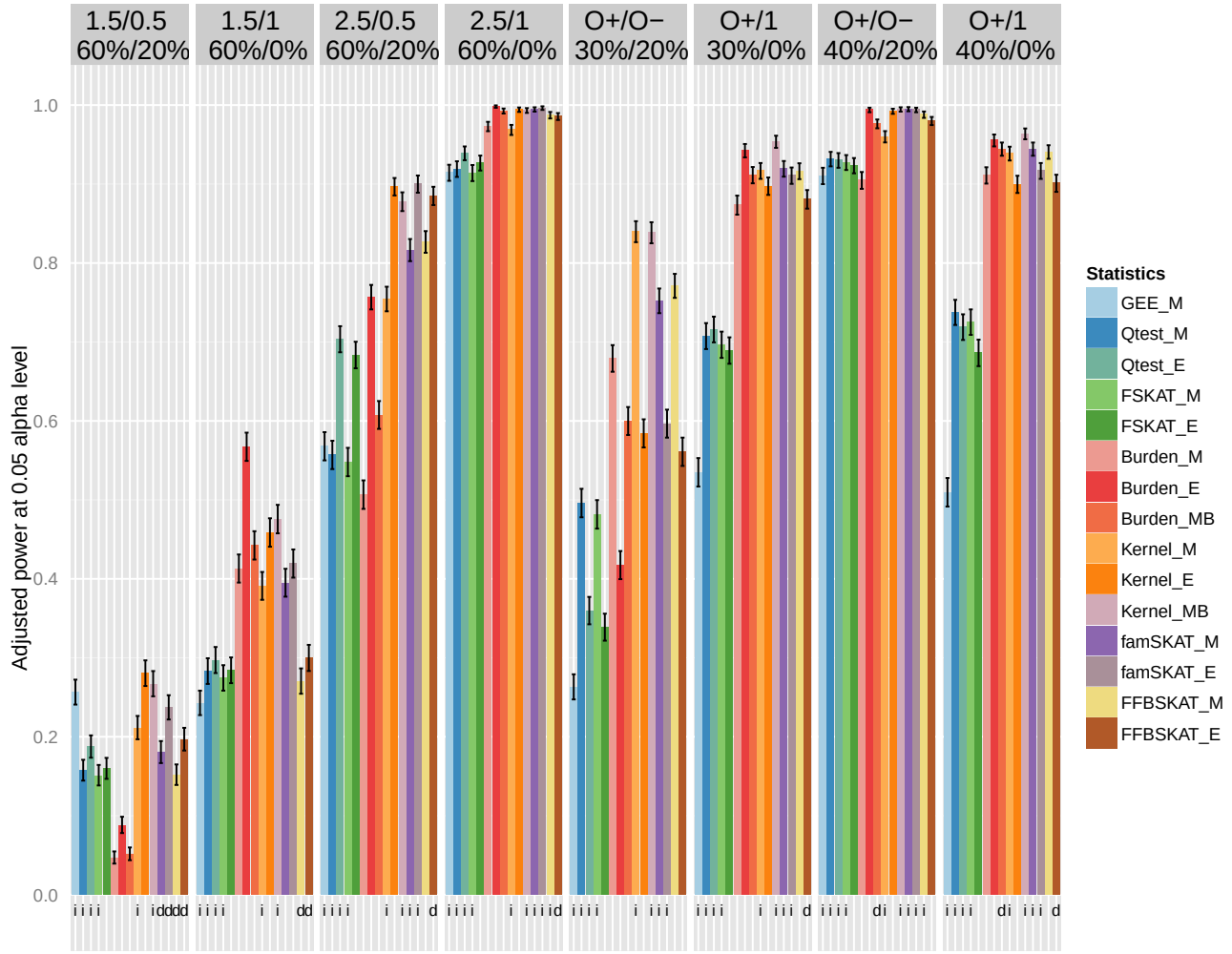


Figure A2: Adjusted power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.05 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights.

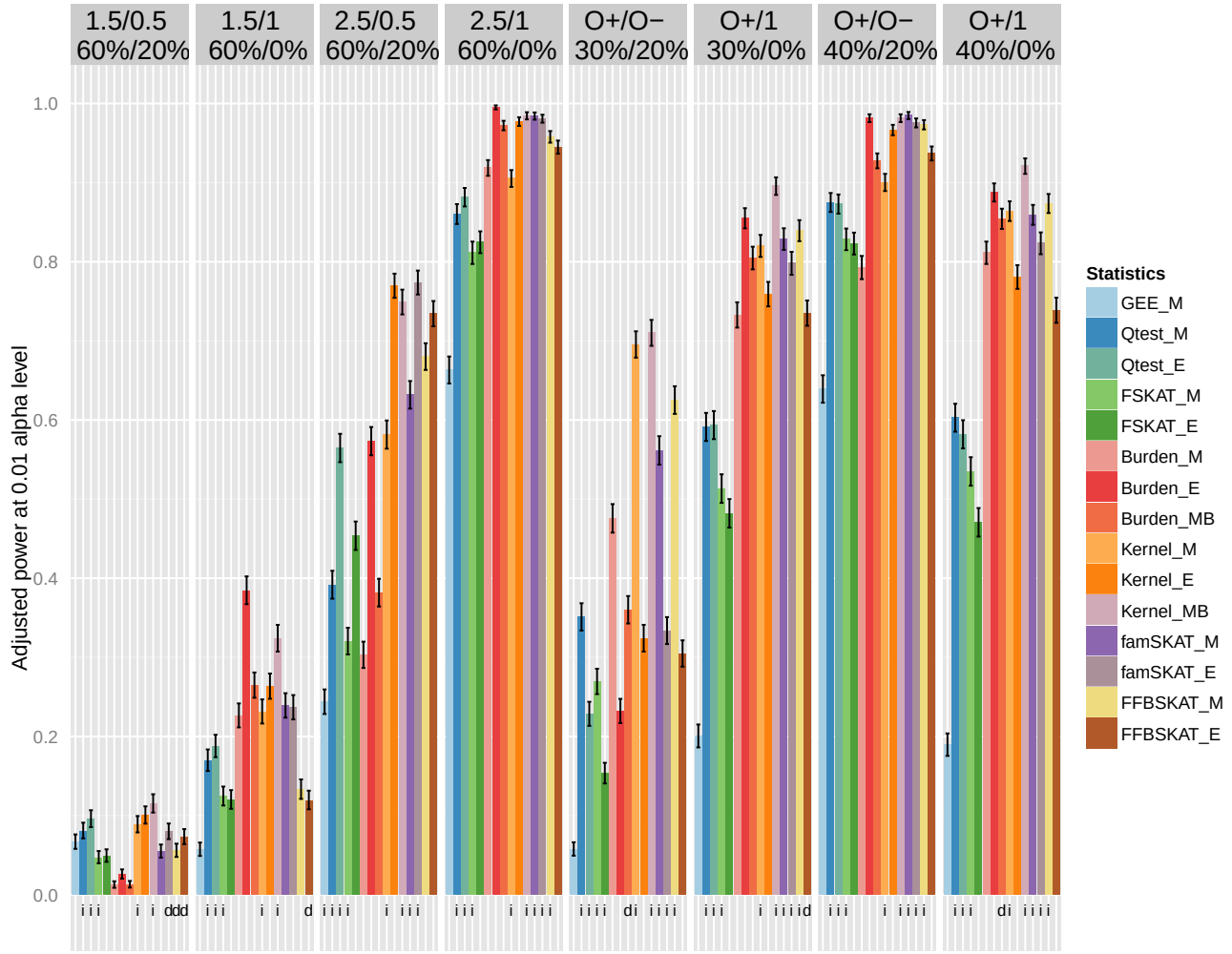


Figure A3: Adjusted power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.01 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights.

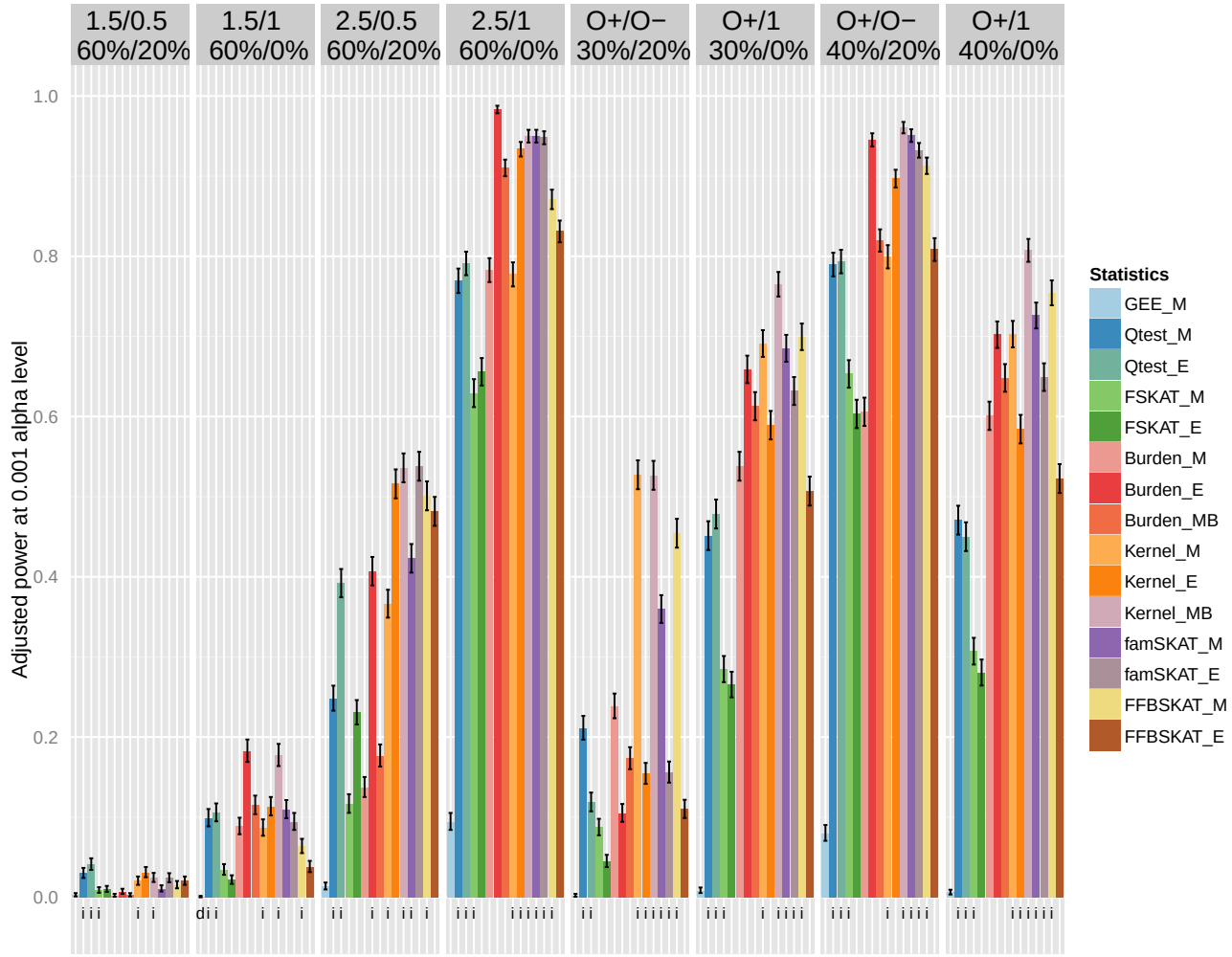
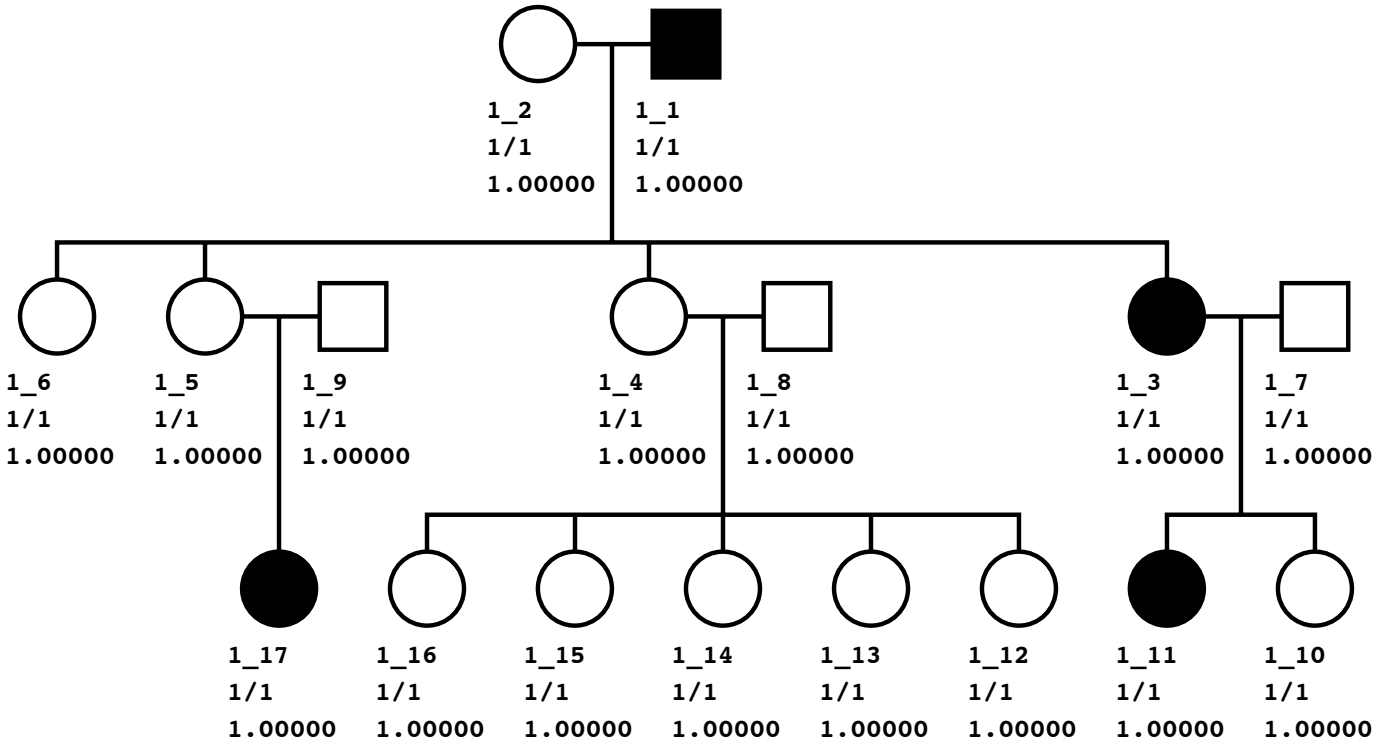


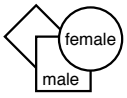
Figure A4: Adjusted power under eight scenarios (ordered as scenarios 1 - 8 from left to right) at the 0.001 alpha level. Odds Ratio (Risk/Protective), Percentage (Risk/Protective). $O^+ = exp^{\frac{ln(10)}{4}|log_{10}MAF_j|}$, $O^- = exp^{-\frac{ln(10)}{4}|log_{10}MAF_j|}$, MAF_j: minor allele frequency for the jth marker calculated in haplotype pool. Bottom Labels: "i": Inflated Type I error, "d": Deflated Type I error. "M": sample-MAF-dependent weights; "E": equal weights; "MB": Madsen-Browning weights.

APPENDIX B

SIMULATED FAMILIES FOR RARE VARIANT ASSOCIATION ANALYSIS

17 individuals





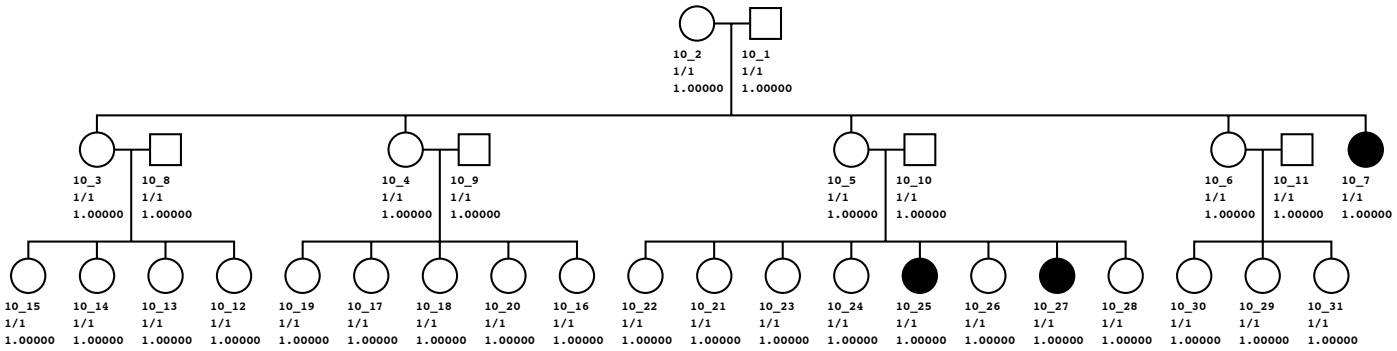
Name
M8
Genotyping-rate

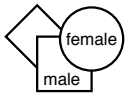
Affected

Unknown

Trait

31 individuals





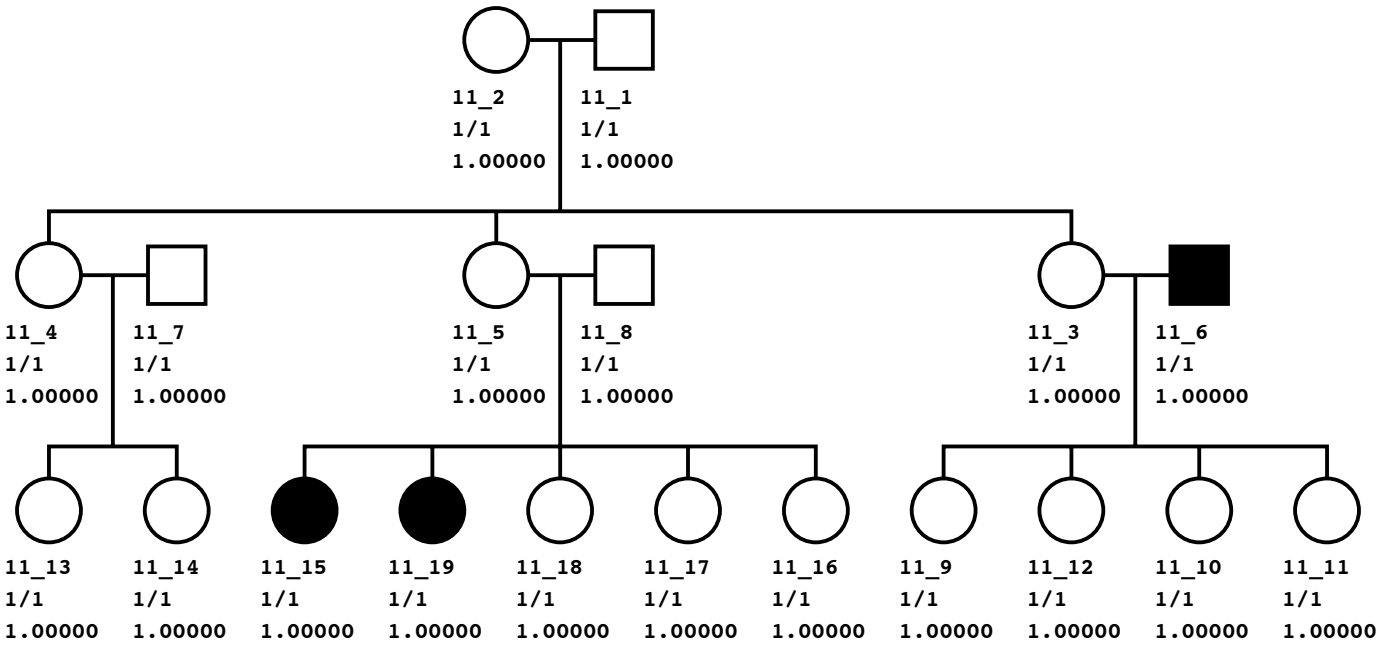
Name
M8
Genotyping-rate

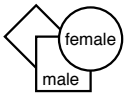
Affected

Unknown

Trait

19 individuals





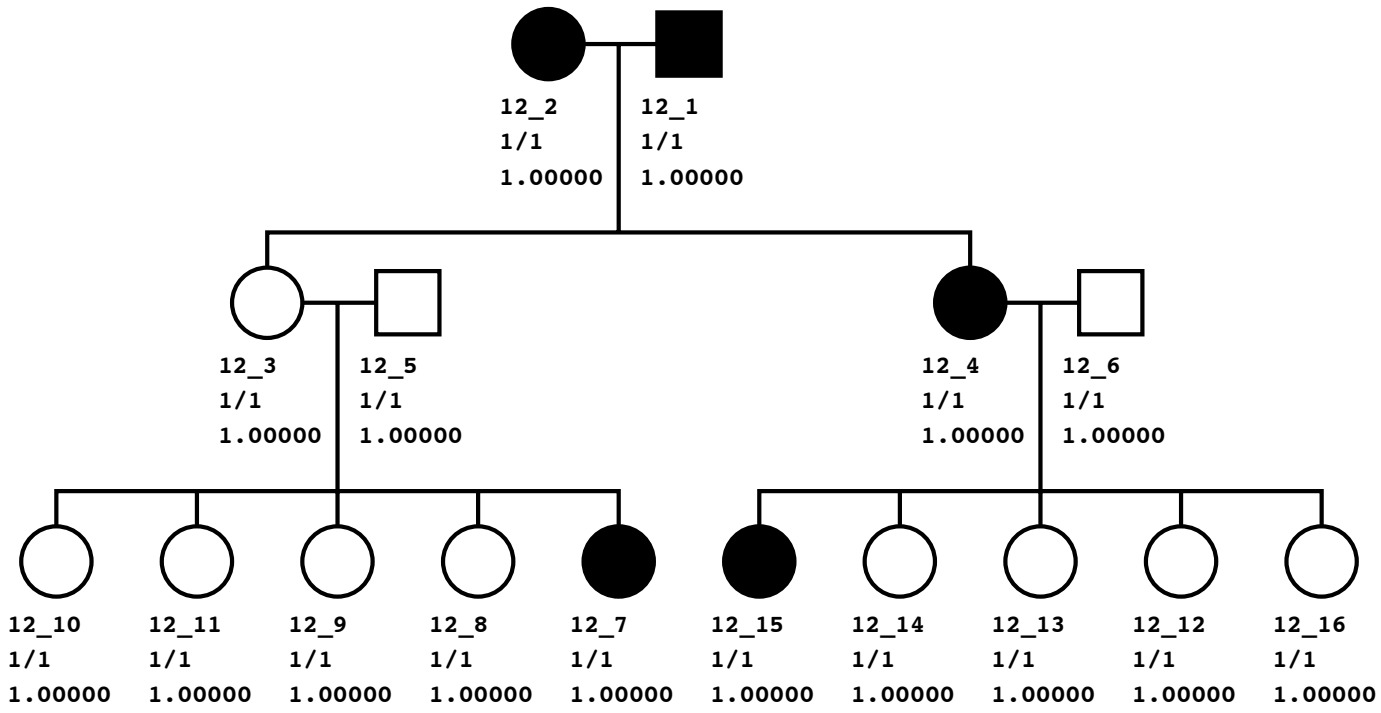
Name
M8
Genotyping-rate

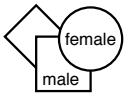
Affected

Unknown

Trait

16 individuals





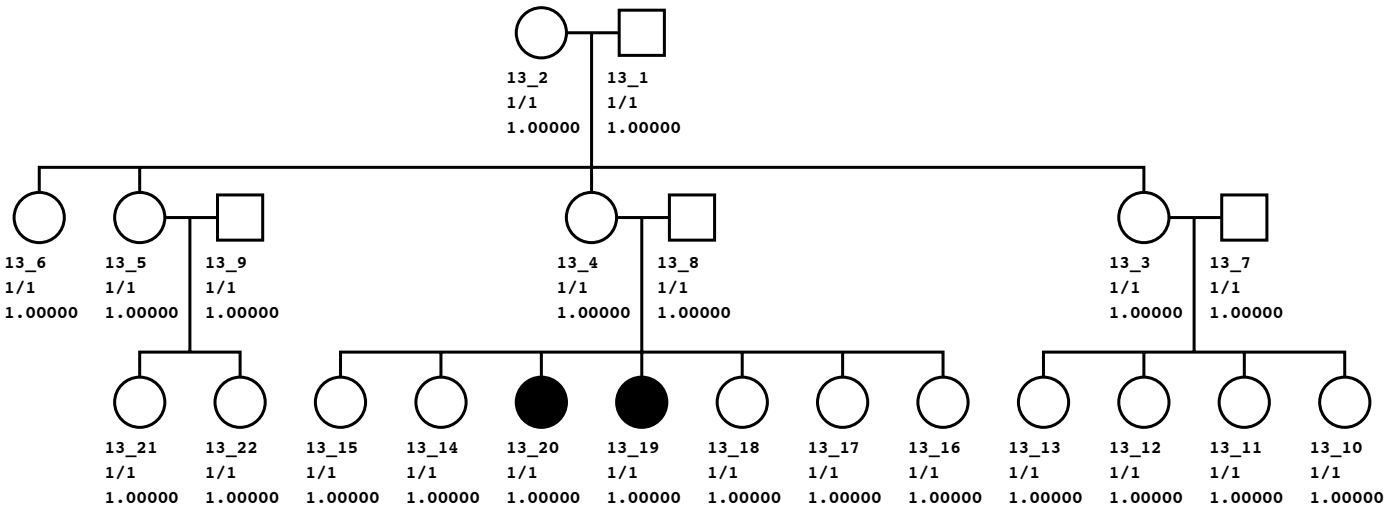
Name
M8
Genotyping-rate

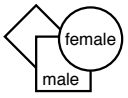
Affected

Unknown

Trait

22 individuals





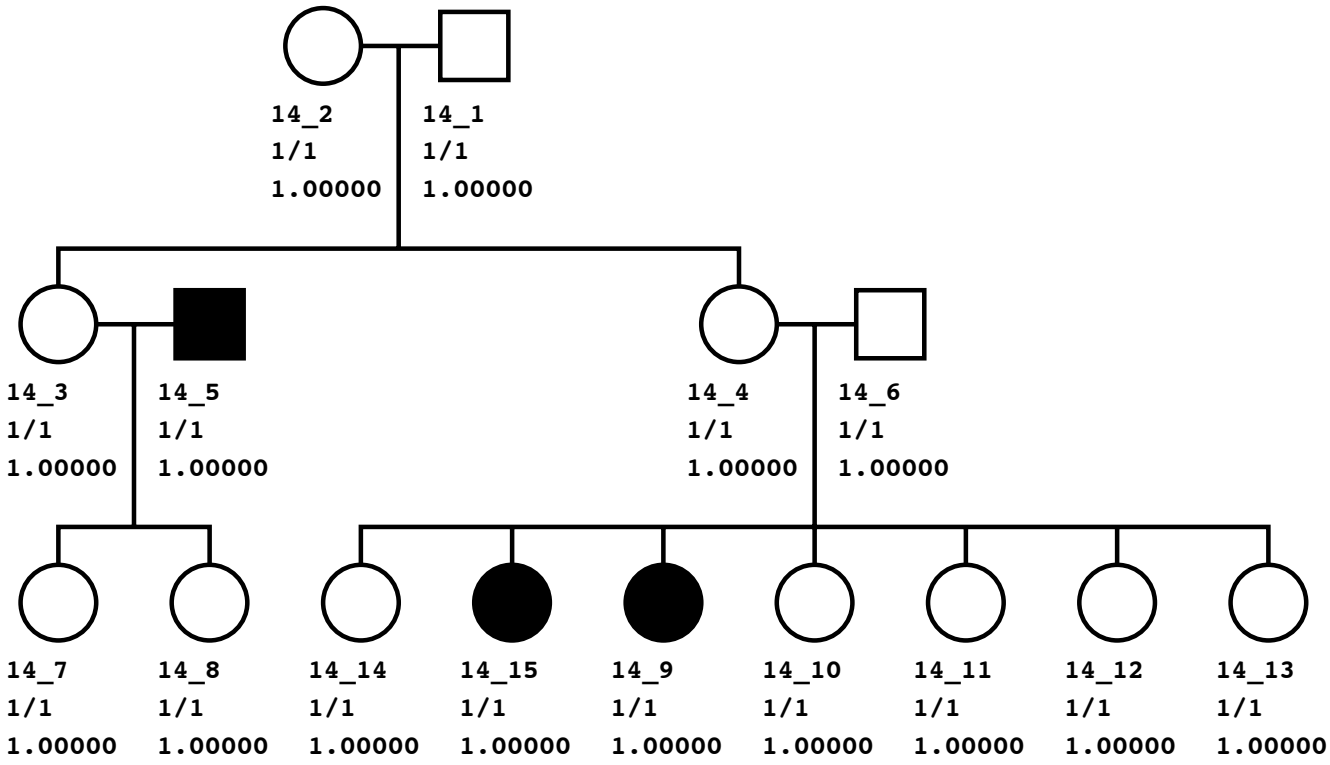
Name
M8
Genotyping-rate

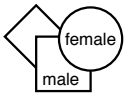
Affected

Unknown

Trait

15 individuals





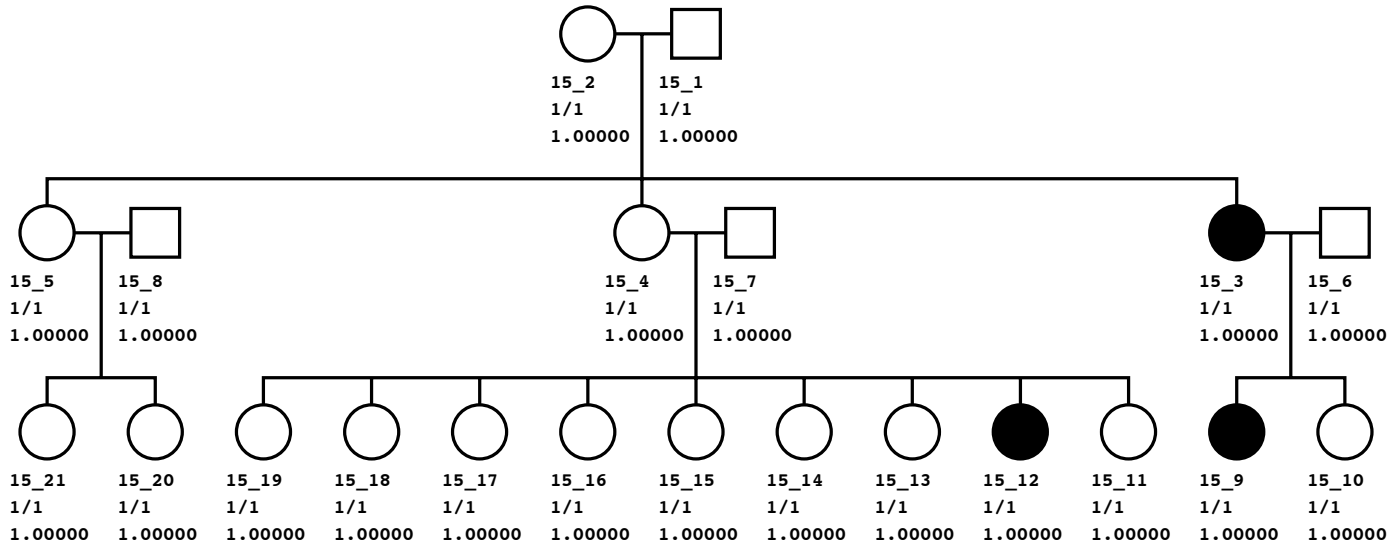
Name
M8
Genotyping-rate

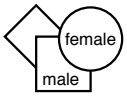
Affected

Unknown

Trait

21 individuals





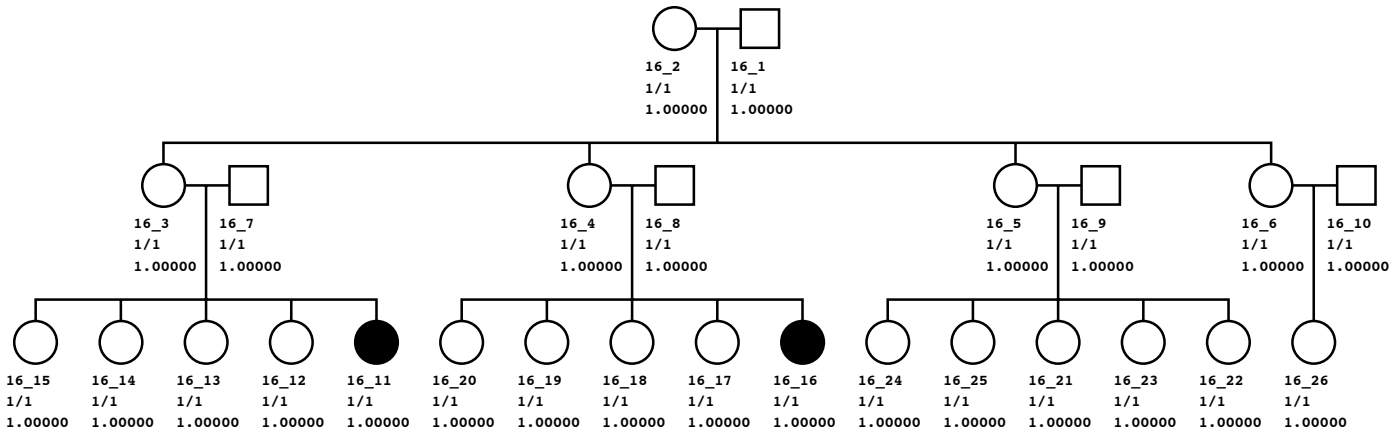
Name
M8
Genotyping-rate

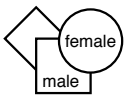
Affected

Unknown

Trait

26 individuals





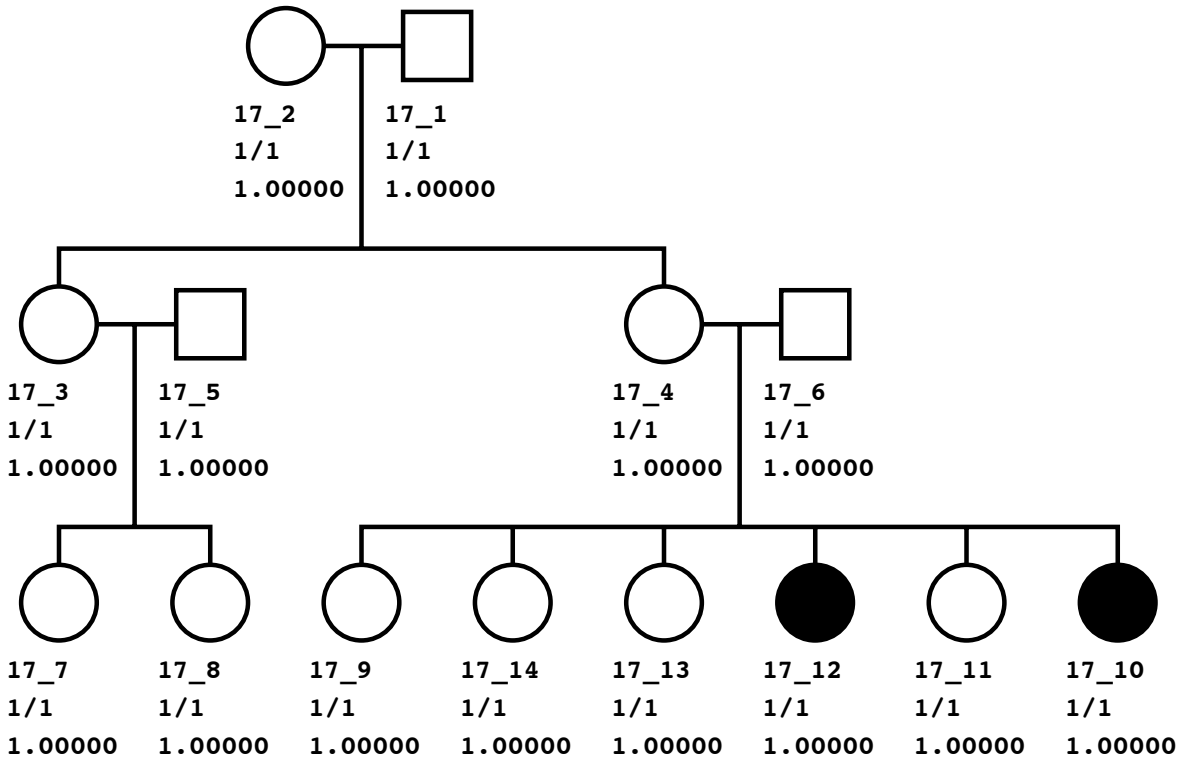
Name
M8
Genotyping-rate

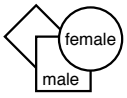
Affected

Unknown

Trait

14 individuals



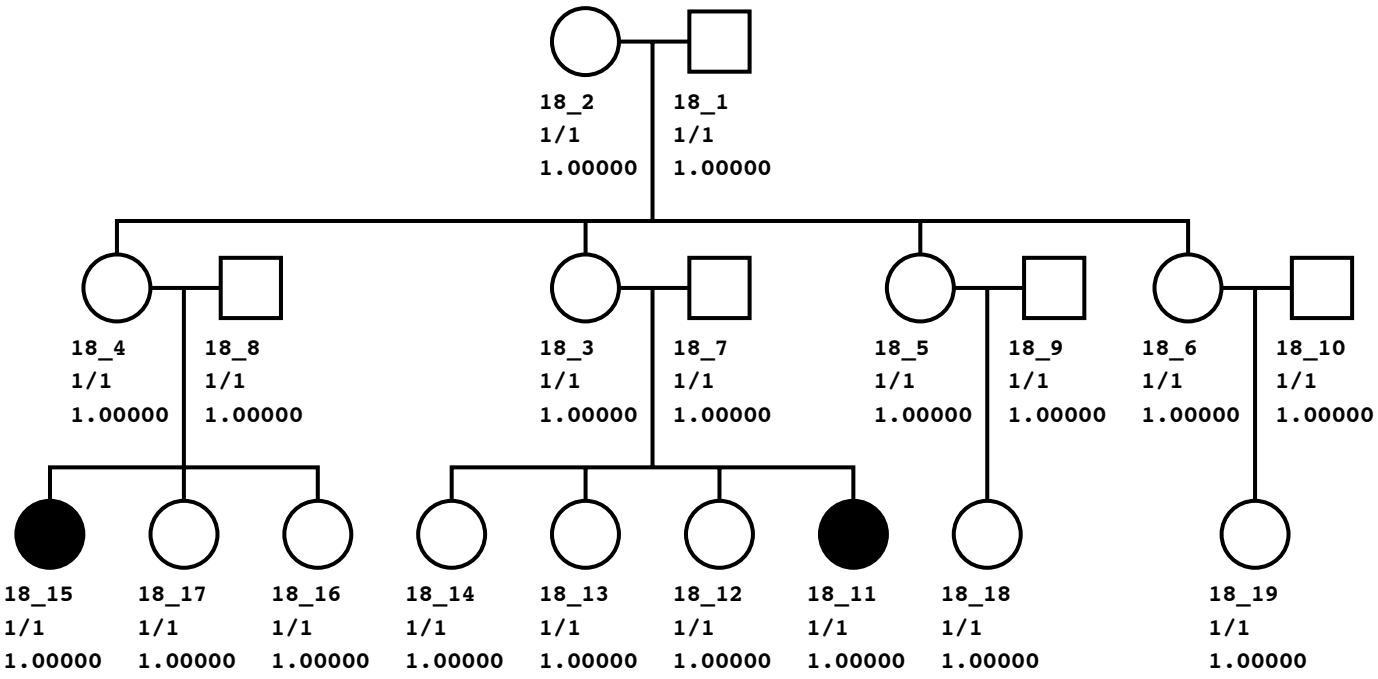


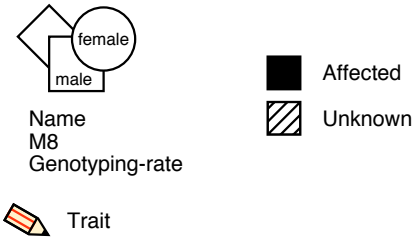
Name
M8
Genotyping-rate

Affected

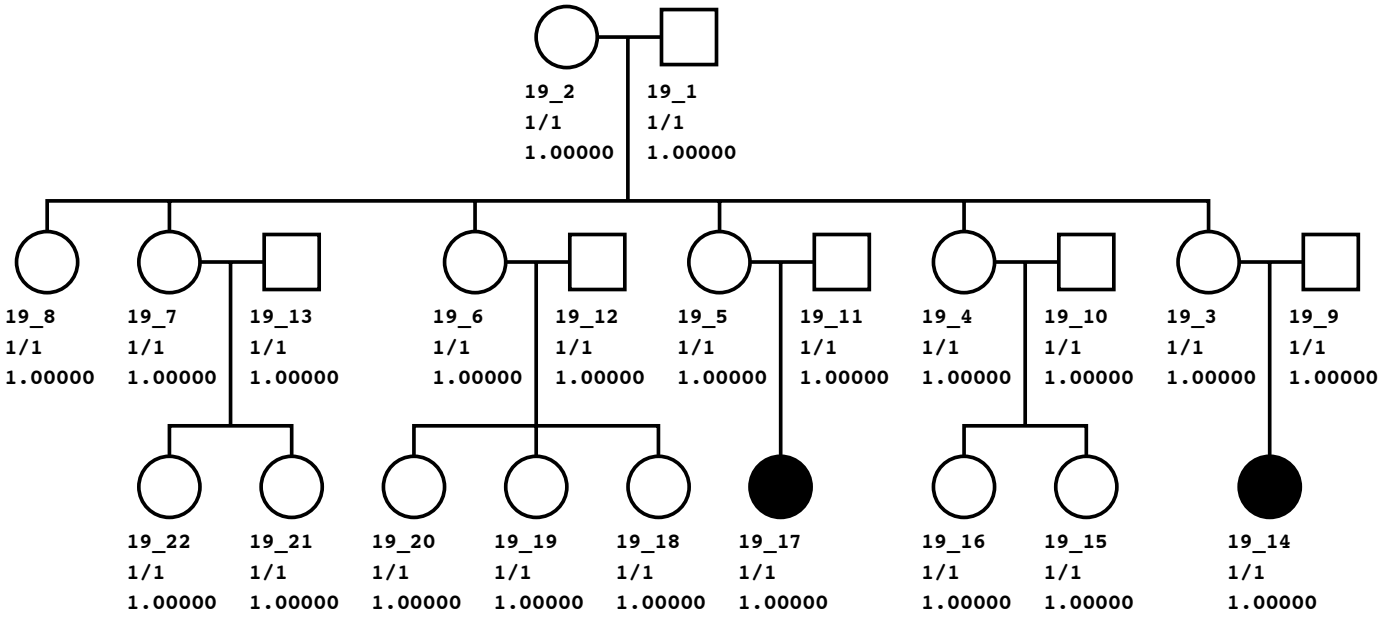
Unknown

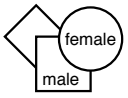
Trait


18_2
1/1
1.00000
18_1
1/1
1.00000
18_4
1/1
1.00000
18_8
1/1
1.00000
18_3
1/1
1.00000
18_7
1/1
1.00000
18_5
1/1
1.00000
18_9
1/1
1.00000
18_6
1/1
1.00000
18_10
1/1
1.00000
18_15
1/1
1.00000
18_17
1/1
1.00000
18_16
1/1
1.00000
18_14
1/1
1.00000
18_13
1/1
1.00000
18_12
1/1
1.00000
18_11
1/1
1.00000
18_18
1/1
1.00000
18_19
1/1
1.00000


female
male
Name
M8
Genotyping-rate
Affected
Unknown
Trait

22 individuals





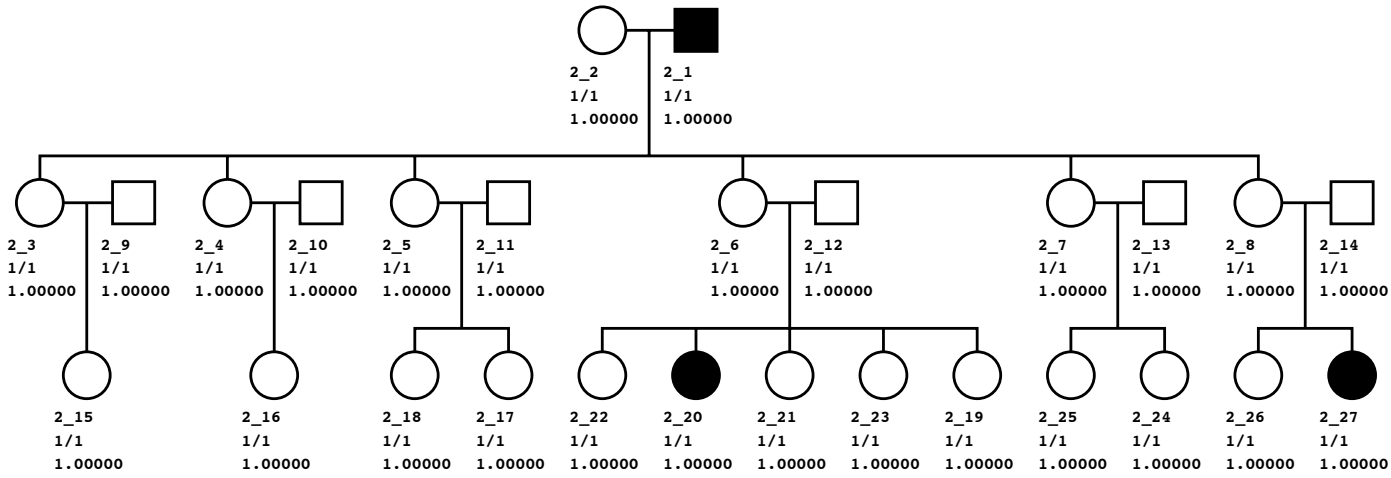
Name
M8
Genotyping-rate

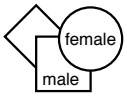
Affected

Unknown

Trait

27 individuals





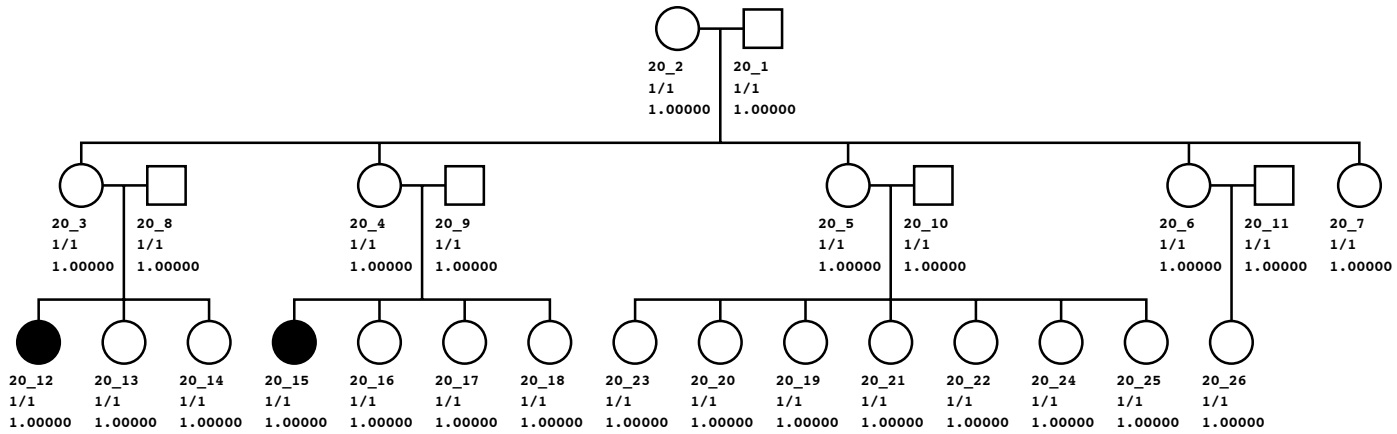
Name
M8
Genotyping-rate

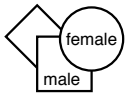
Affected

Unknown

Trait

26 individuals





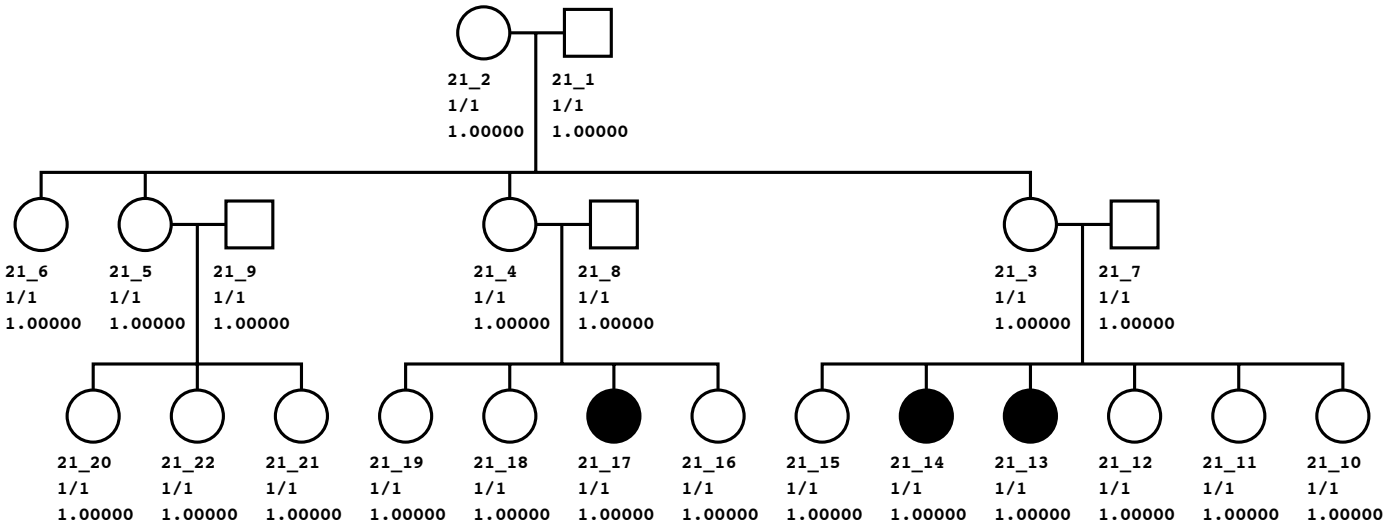
Name
M8
Genotyping-rate

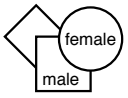
Affected

Unknown

Trait

22 individuals





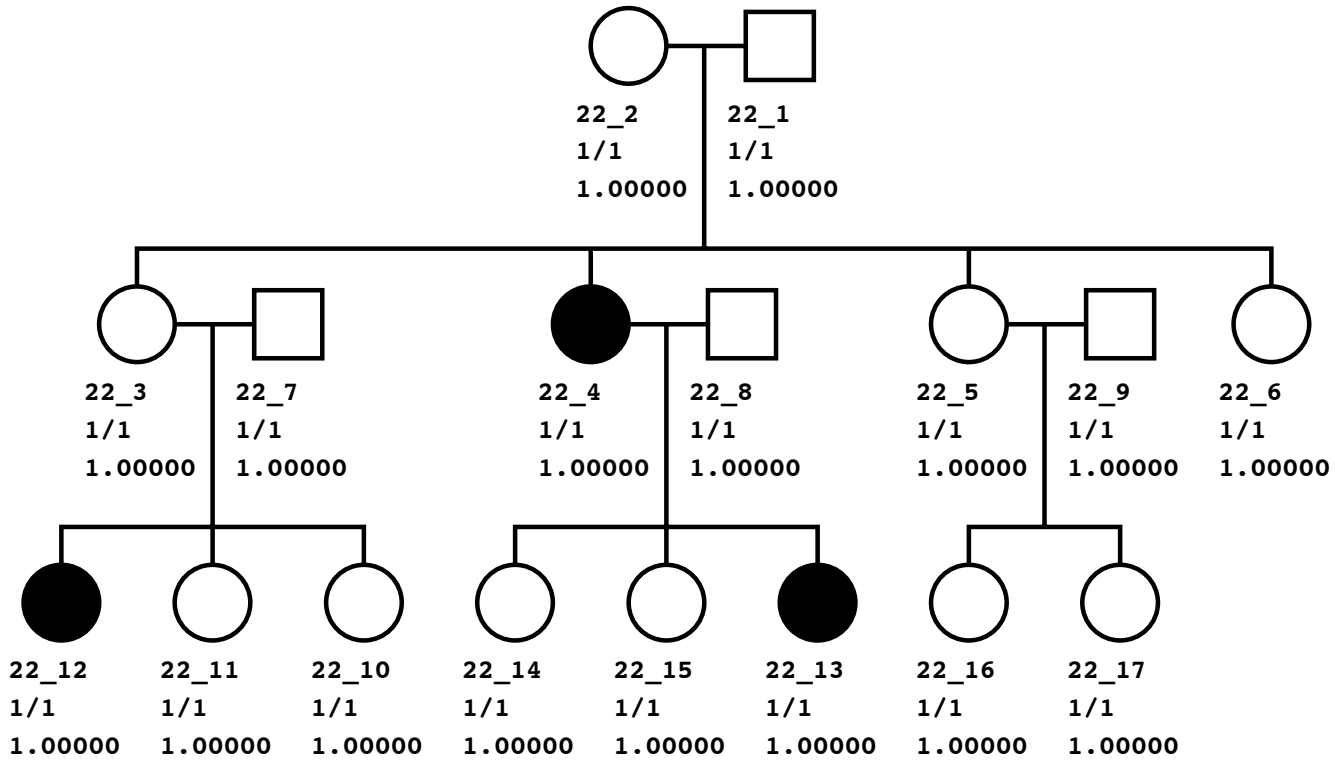
Name
M8
Genotyping-rate

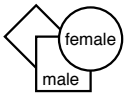
Affected

Unknown

Trait

17 individuals





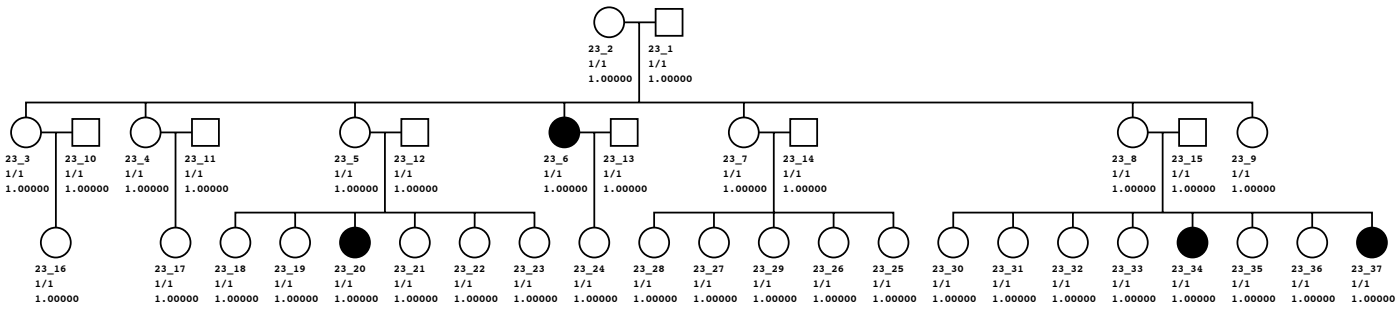
Name
M8
Genotyping-rate

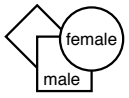
Affected

Unknown

Trait

37 individuals





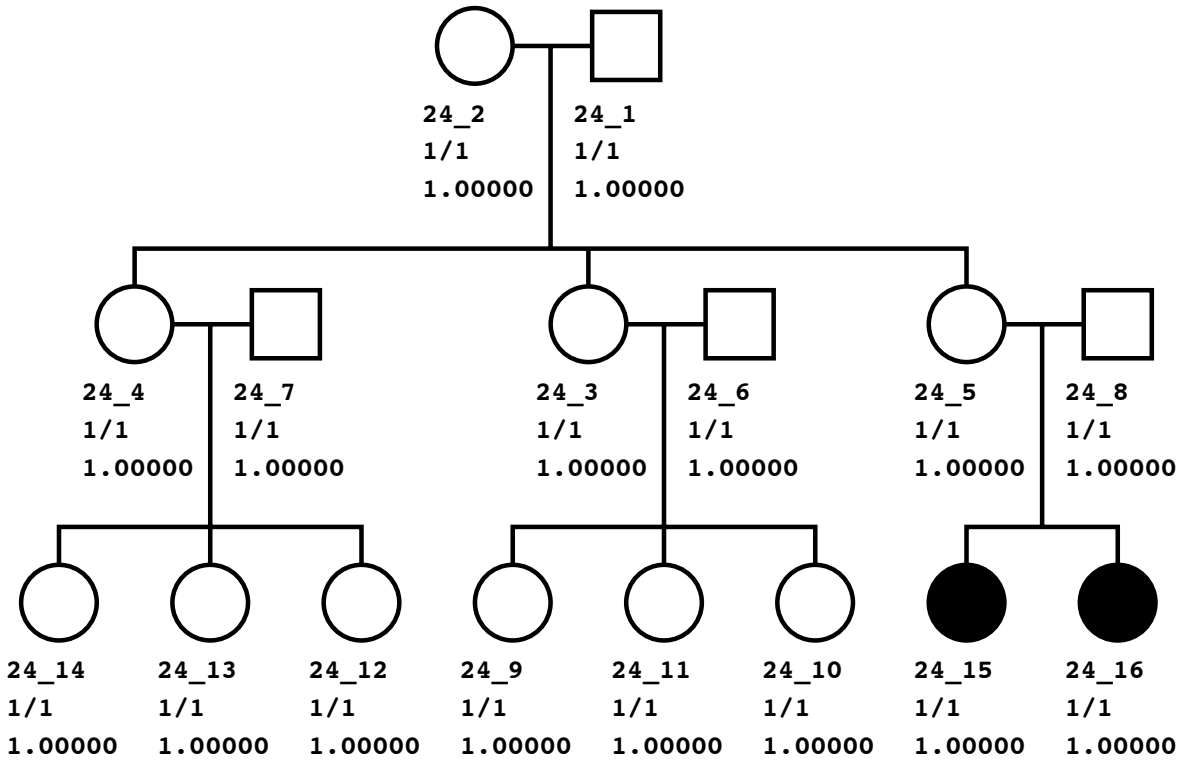
Name
M8
Genotyping-rate

Affected

Unknown

Trait

16 individuals

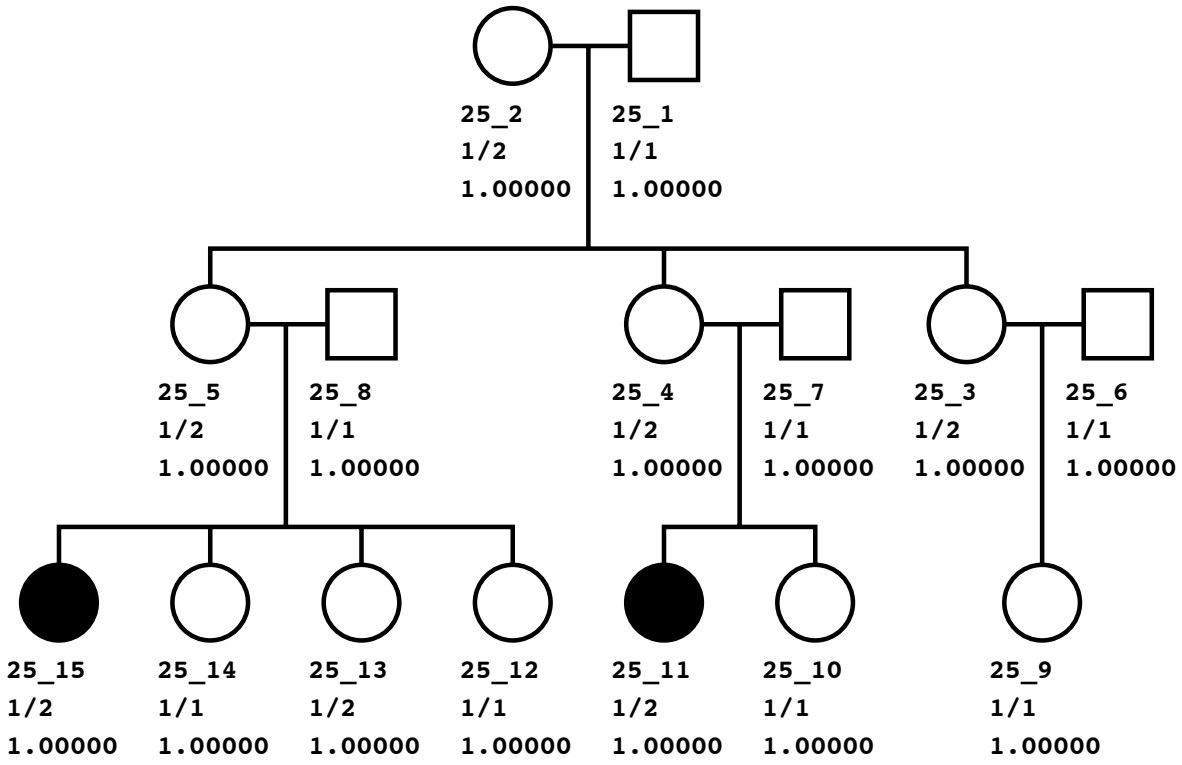


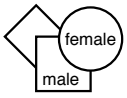
female
male
Name
M8
Genotyping-rate

Affected
Unknown

Trait

15 individuals





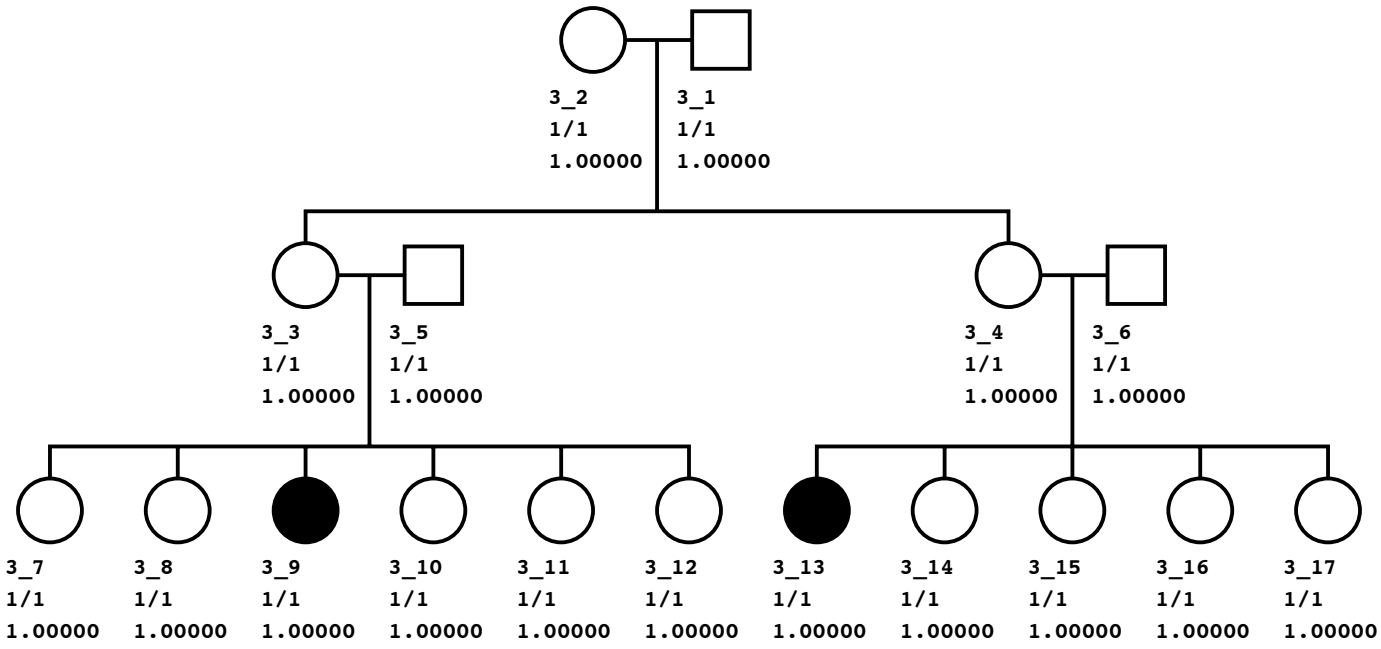
Name
M8
Genotyping-rate

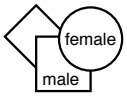
Affected

Unknown

Trait

17 individuals





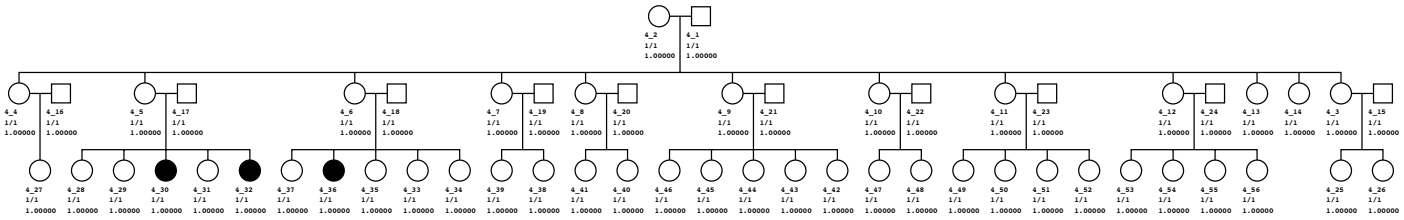
Name
M8
Genotyping-rate

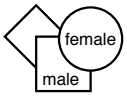
Affected

Unknown

Trait

56 individuals





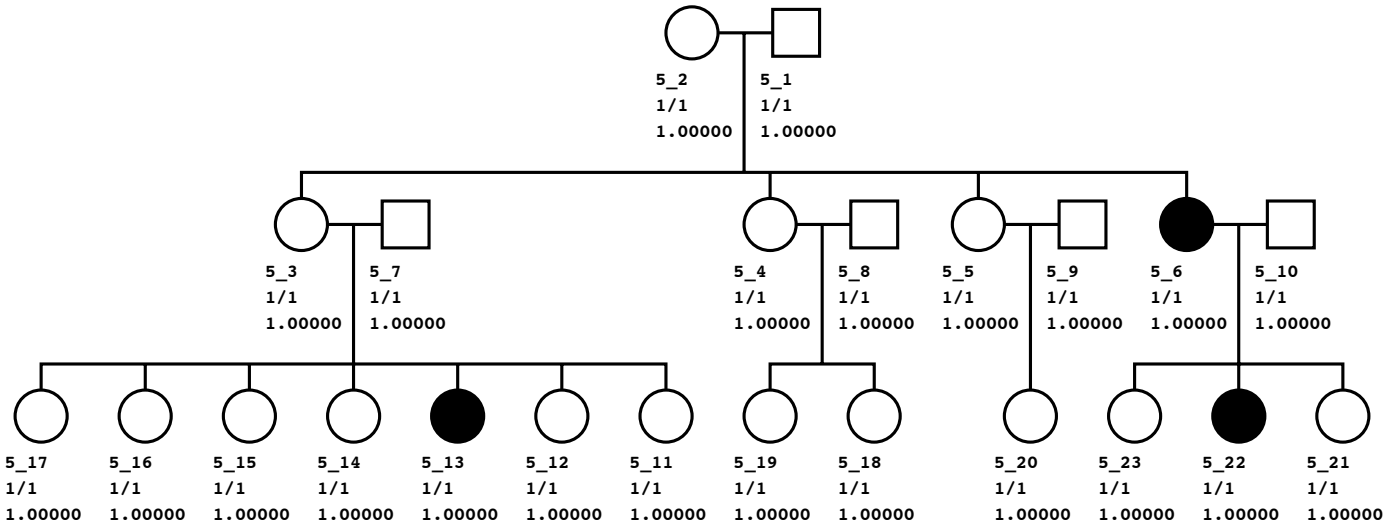
Name
M8
Genotyping-rate

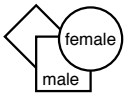
Affected

Unknown

Trait

23 individuals





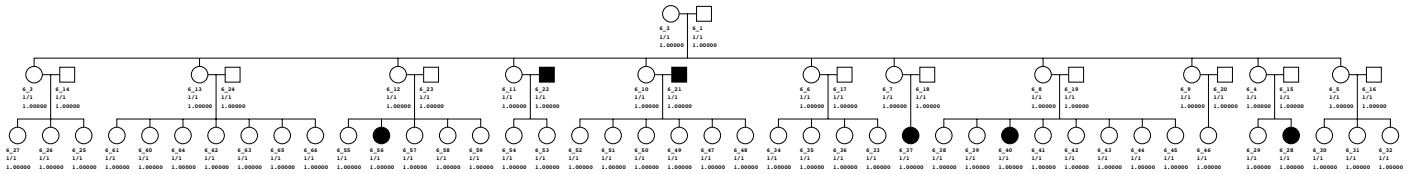
Name
M8
Genotyping-rate

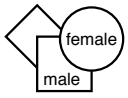
Affected

Unknown

Trait

66 individuals





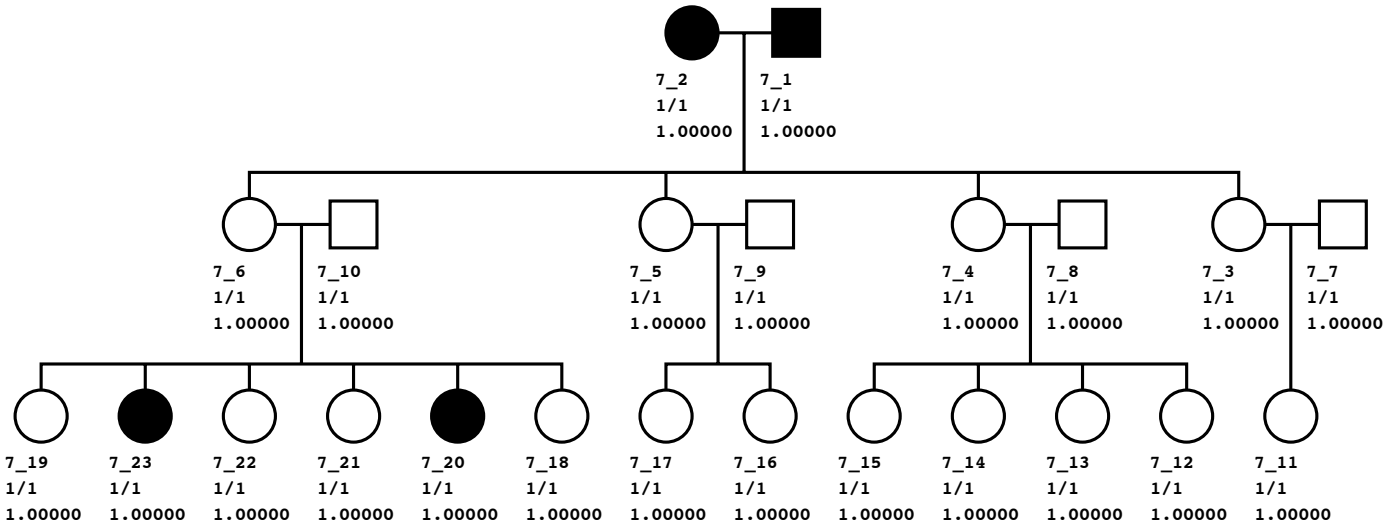
Affected

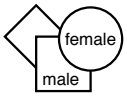
Unknown

Name
M8
Genotyping-rate

Trait

23 individuals





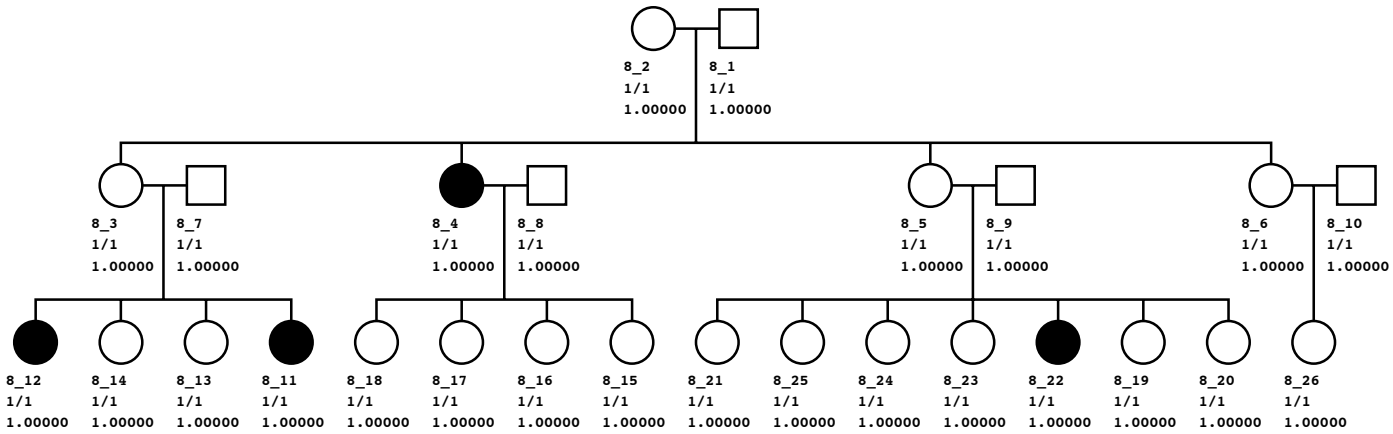
Name
M8
Genotyping-rate

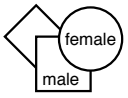
Affected

Unknown

Trait

26 individuals





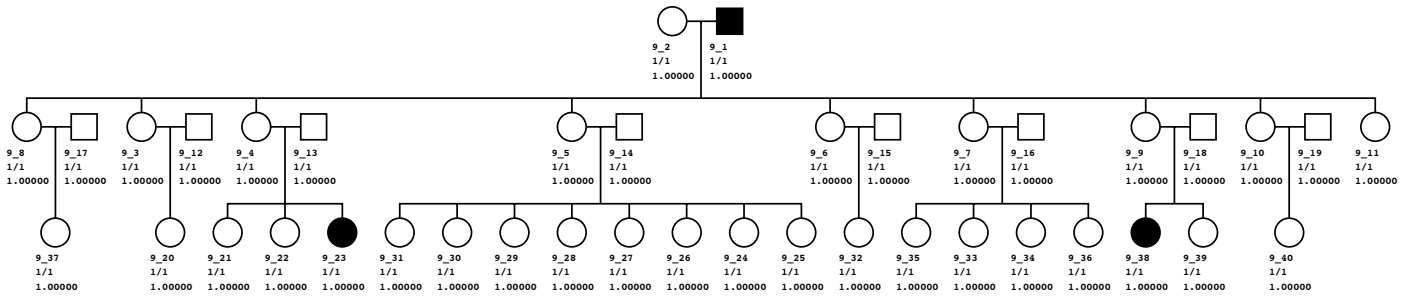
Name
M8
Genotyping-rate

Affected

Unknown

Trait

40 individuals



female
male

Name
M8
Genotyping-rate

Affected

Unknown

Trait

APPENDIX C

R IMPLEMENTATION OF THE QTEST

```
Qtest <- function(phenoQ, genoQ, id, fa, mo, family="binomial", weights=NULL, covariates=NULL){
  # phenoQ: A vector contains phenotypes for all the subjects, no missing.
  # genoQ: A vector or matrix c ontains genotypes for all the subjects, no missing.
  # id: A vector contains subjects' IDs, no missing.
  # fa: A vector contains fathers' IDs, should be unique in the sample, no missing.
  # mo: A vector contains mothers' IDs, should be unique in the sample, no missing.
  # family: optional, specify the distribution of the trait, default is "binomial".
  # weights: optional, 1) "Equal" means equal weights, 2) default is sample-MAF-dependent, or 3) an vector contains user specified weights for the marker(s).
  # covariates: optinal, 1) default is NULL, 2) an vector of matrix contains covariates of interest.
  library(MASS)
  library(kinship2)
  library(nlme)
  library(mgcv)
  if(class(phenoQ) != "data.frame") stop("phenoQ should be data.frame class!")
  if(class(genoQ) != "data.frame") stop("genoQ should be data.frame class!")
  n1 <- nrow(phenoQ)
  n2 <- nrow(genoQ)
  if(n1 != n2) stop("Number of subjects in phenoQ and genoQ files do not match")
  y <- as.matrix(phenoQ)
  K <- kinship(id, fa, mo)
  if (weights=="Equal"){
    W <- diag(1, n2)
  } else {
    w <- dbeta(colMeans(genoQ)/2, 1, 25)
    W <- diag(w^2)
  }
  intercept <- rep(1,length(id))
  if(is.null(covariates)){
    X <- as.matrix(intercept)
    exprs<-paste("y ~ 1")} else if(!is.null(covariates)){
      X <- cbind(intercept, as.matrix(covariates))
      X <- as.matrix(X)
      exprs<-paste("y ~", paste(names(covariates),collapse=" + "))}
  cs.K <- corSymm(2*K[lower.tri(K)],fixed=T)
  id <- as.matrix(id)
  colnames(id) <- "id"
  cs.K <- Initialize(cs.K, data = id)
  data <- data.frame(id = as.factor(id), y = y)
  fit1 <- glmmPQL2(as.formula(exprs), random = ~1|id, correlation = cs.K, data = data, family = family, control = lmeControl(opt = "optim"))
  G <- as.matrix(genoQ)
```

```
  II <- diag(1, nrow(G))
  alpha <- fit1$fit$coefficients$fixed
  V_beta <- extract.lme.cov(fit1$fit, data) # '1' -- calculated from this function
  V_beta_inv <- solve(V_beta)
  lamda <- V_beta_inv - V_beta_inv %*% X %*% solve(t(X) %*% V_beta_inv %*% X) %*% t(X) %*% V
      _beta_inv
  res <- as.matrix(Y <- fit1$y_star - X %*% alpha)
  KK <- 2 * K
  T_11 <- 1/2 * sum(diag(lamda %*% G %*% W %*% t(G) %*% lamda %*% G %*% W %*% t(G)))
  T_13 <- 1/2 * sum(diag(lamda %*% G %*% W %*% t(G) %*% lamda %*% II %*% KK %*% t(II)))
  T_31 <- t(T_13)
  T_33 <- 1/2 * sum(diag(lamda %*% II %*% KK %*% t(II) %*% lamda %*% II %*% KK %*% t(II)))
  I <- T_11 - T_13 * T_31/T_33
  e <- 1/2 * sum(diag(lamda %*% G %*% W %*% t(G)))
  k <- I/(2*e)
  v <- 2*e^2/I
  U <- 1/2 * (t(res) %*% V_beta_inv %*% G %*% W %*% t(G) %*% V_beta_inv %*% (res))
  S <- U/k
  Q <- pchisq(S, df = v, lower.tail = FALSE)
  result <- round(Q, digits = 5)
  return(result)
}
########################################
glmmPQL2 <- function (fixed, random, family, data, correlation, weights,
                      control, niter = 10, verbose = TRUE, ...)
{# Modified glmmPQL function, which returns the y values at the last iteration.
  if (!require("nlme"))
    stop("package 'nlme' is essential")
  if (is.character(family))
    family <- get(family)
  if (is.function(family))
    family <- family()
  if (is.null(family$family)) {
    print(family)
    stop("'family' not recognized")
  }
  m <- mcall <- Call <- match.call()
  nm <- names(m)[-1L]
  keep <- is.element(nm, c("weights", "data", "subset", "na.action"))
  for (i in nm[!keep]) m[[i]] <- NULL
  allvars <- if (is.list(random))
    allvars <- c(all.vars(fixed), names(random), unlist(lapply(random, function(x) all.vars(
        formula(x)))))
  else c(all.vars(fixed), all.vars(random))
  Terms <- if (missing(data))
    terms(fixed)
  else terms(fixed, data = data)
  off <- attr(Terms, "offset")
  if (length(off <- attr(Terms, "offset")))
    allvars <- c(allvars, as.character(attr(Terms, "variables"))[off + 1])
  if (!missing(correlation) && !is.null(attr(correlation, "formula")))
    allvars <- c(allvars, all.vars(attr(correlation, "formula")))
  Call$fixed <- eval(fixed)
  Call$random <- eval(random)
  m$formula <- as.formula(paste("~", paste(allvars, collapse = "+")))
  environment(m$formula) <- environment(fixed)
  m$drop.unused.levels <- TRUE
  m[[1L]] <- as.name("model.frame")
  mf <- eval.parent(m)
  off <- model.offset(mf)
  if (is.null(off))
    off <- 0
  wts <- model.weights(mf)
  if (is.null(wts))
    wts <- rep(1, nrow(mf))
  mf$wts <- wts
  fit0 <- glm(formula = fixed, family = family, data = mf, weights = wts, ...)
  w <- fit0$prior.weights
```

```
  eta <- fit0$linear.predictors
  zz <- eta + fit0$residuals - off
  wz <- fit0$weights
  fam <- family
  nm <- names(mcall)[-1L]
  keep <- is.element(nm, c("fixed", "random", "data", "subset", "na.action", "control"))
  for (i in nm[!keep]) mcall[[i]] <- NULL
  fixed[[2L]] <- quote(zz)
  mcall[["fixed"]] <- fixed
  mcall[[1L]] <- as.name("lme")
  mcall$random <- random
  mcall$method <- "ML"
  if (!missing(correlation))
    mcall$correlation <- correlation
  mcall$weights <- quote(varFixed(~invwt))
  mf$zz <- zz
  mf$invwt <- 1/wz
  mcall$data <- mf
  for (i in seq_len(niter)) {
    if (verbose)
      message("iteration ", i)
    fit <- eval(mcall)
    etaold <- eta
    eta <- fitted(fit) + off
    if (sum((eta - etaold)^2) < 1e-06 * sum(eta^2))
      break
    mu <- fam$linkinv(eta)
    mu.eta.val <- fam$mu.eta(eta)
    mf$zz <- eta + (fit0$y - mu)/mu.eta.val - off
    wz <- w * mu.eta.val^2/fam$variance(mu)
    mf$invwt <- 1/wz
    mcall$data <- mf
  }
  y_star <- mf$zz
  attributes(fit$logLik) <- NULL
  fit$call <- Call
  fit$family <- family
  fit$logLik <- as.numeric(NA)
  oldClass(fit) <- c("glmmPQL", oldClass(fit))
  newfit <- list("fit"=fit, "y_star"=y_star, "W_inverse"=mf$invwt)
}
```

Qtest_and_glmmPQL2.R

BIBLIOGRAPHY

G. R. Abecasis and J. E. Wigginton. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am. J. Hum. Genet*, 77:754–767, 2005.

G. R. Abecasis, L. R. Cardon, and W. O. C. Cookson. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet*, 66(1):279–292, 2000a.

G. R. Abecasis, W. O. Cookson, and L. R. Cardon. Pedigree tests of transmission disequilibrium. *Eur J Hum Genet*, 8(7):545–551, 2000b.

G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30(1):97–101, 2002.

Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. van Duijn. Genabel: an r library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–6, 2007.

M. C. Babron, M. de Tayrac, D. N. Rutledge, E. Zeggini, and E. Genin. Rare and low frequency variant stratification in the uk population: description and impact on association tests. *PLoS One*, 7(10):e46519, 2012.

M. Boehnke. Allele frequency estimation from data on relatives. *Am J Hum Genet*, 48(1): 22–25, 1991.

J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285, 1999.

C. Bourgain, S. Hoffjan, R. Nicolae, D. Newman, L. Steiner, K. Walker, R. Reynolds, C. Ober, and M. S. McPeek. Novel case-control test in a founder population identifies p-selectin as an atopy-susceptibility locus. *Am J Hum Genet*, 73(3):612–626, 2003.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

S. R. Browning. Multilocus association mapping using variable-length markov chains. *Am J Hum Genet*, 78:903–913, 2006.

R. M. Cantor, G. K. Chen, P. Pajukanta, and K. Lange. Association testing in a linked region using large pedigrees. *Am J Hum Genet*, 76(3):538–542, 2005.

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

M. Chen, X. Liu, F. Wei, M. G. Larson, C. S. Fox, R. S. Vasan, and Q. Yang. A comparison of strategies for analyszing dichotomous outcomes in genome-wide association studies with general pedigrees. *Genet Epidemiol*, 35(7):650–657, 2011.

W. M. Chen and G. R. Abecasis. Family-based association tests for genomewide association scans. *Am J Hum Genet*, 81(5):913–26, 2007.

W. M. Chen, A. Manichaikul, and S. S. Rich. A generalized family-based association test for dichotomous traits. *Am J Hum Genet*, 85(3):364–76, 2009.

K. F. Cheng and J. H. Chen. Detecting rare variants in case-parents association studies. *PLoS One*, 8(9):e74310, 2013.

C. Y. Cheung, E. A. Thompson, and E. M. Wijsman. Detection of mendelian consistent genotyping errors in pedigrees. *Genet Epidemiol*, 38(4):291–299, 2014.

C. Y. K. Cheung, E. A. Thompson, and E. M. Wijsman. Gigi: An approach to effective imputation of dense genotypes on large pedigrees. *American Journal of Human Genetics*, 92:504–516, 2013.

Y. H. Cheung, G. Wang, S. M. Leal, and S. Wang. A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. *Genet Epidemiol*, 36(7):675–85, 2012.

A. G. Clark, M. J. Hubisz, C. D. Bustamante, S. H. Willianmson, and R. Nielsen. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*, 15(11): 1496–1502, 2005.

D. Clayton. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet*, 65(4):1170–7, 1999.

D. Clayton. Testing for association on the x chromosome. *Biostatistics*, 9(4):593–600, 2008.

Jr. Cottingham, R. W., R. M. Idury, and A. A. Schaffer. Faster sequential genetic linkage computations. *Am J Hum Genet*, 53(1):252–63, 1993.

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

M De Andrade and CI Amos. Ascertainment issues in variance components models. *Genet Epidemiol*, 19:333–344, 2000.

F. Dudbridge. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered*, 66(2):87–98, 2008.

R. C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Hum Hered*, 21:523–542, 1971.

H. Fang, B. Hou, Q. Wang, and Y. Yang. Rare variants analysis by risk-based variable-threshold method. *Comput Biol Chem*, 46:32–8, 2013.

D. W. Fulker, S. S. Cherny, P. C. Sham, and J. K. Hewitt. Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet*, 64(1):259–267, 1999.

H. H. Göring and J. D. Terwilliger. Linkage analysis in the presence of errors iv: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet*, 66(4):1310–1327, 2000.

Z. He, B. J. O'Roak, J. D. Smith, G. Wang, S. Hooker, R. L. Santos-Cortez, B. Li, M. Kan, N. Krumm, D. A. Nickerson, J. Shendure, E. E. Eichler, and S. M. Leal. Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am J Hum Genet*, 94(1):33–46, 2014.

S. C. Heath. Markov chain monte carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet*, 61:748–760, 1997.

T. Hiekkalinna, A. A. Schaffer, B. Lambert, P. Norrgrann, H. H. Göring, and J. D. Terwilliger. Pseudomarker: a powerful program for joint linkage and/or linkage disequilibrium analysis on mixtures of singletons and related individuals. *Hum Hered*, 71(4):256–66, 2011.

T. Hiekkalinna, H. H. Göring, B. Lambert, K. M. Weiss, P. Norrgrann, A. A. Schaffer, and J. D. Terwilliger. On the statistical properties of family-based association tests in datasets containing both pedigrees and unrelated case-control samples. *Eur J Hum Genet*, 20(1476-5438 (Electronic)):217–223, 2012.

M. Huang and D. Zhang. Testing polynomial covariate effects in linear and generalized linear mixed models. *Stat Surv*, 2:154–169, 2008.

I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, and X. Lin. Family-based association tests for sequence data, and comparisons with population-based association tests. *European Journal of Human Genetics*, 21(10):1158–162, Oct 2013.

S. K. Iyengar and R. C. Elston. The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods Mol Biol*, 376:71–84, 2007.

D. Jiang and M. S. McPeek. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet Epidemiol*, 38(1):10–20, 2014.

Y. Jiang, M. P. Epstein, and K. N. Conneely. Assessing the impact of population stratification on association studies of rare variation. *Hum Hered*, 76(1):28–35, 2013.

I. R. Konig, C. Loley, J. Erdmann, and A. Ziegler. How to include chromosome x in your genome-wide association study. *Genetic Epidemiology*, 38(2):97–103, Feb 2014.

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

N. M. Laird, S. Horvath, and X. Xu. Implementing a unified approach to family-based tests of association. *Genet Epidemiol*, 19 Suppl 1(S1):S36–S42, 2000.

E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci.*, 84:2263–2267, 1987.

K. Lange and E. Sobel. Descent graphs in pedigree analysis: applications to haplotyping, locationscores, and marker-sharing statistics. *Am. J. Hum. Genet*, 58:1323–1337, 1996.

K. Lange, R. Cantor, S. Horvath, M. Perola, C. Sabatti, J. Sinsheimer, and E. Sobel. Mendel version 4.0: a complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am J Hum Genet*, 69(4):504–504, 2001.

K. Lange, J. S. Sinsheimer, and E. Sobel. Association testing with mendel. *Genet Epidemiol*, 29(1):36–50, 2005.

L. C. Lazzeroni and K. Lange. A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered*, 48(2):67–81, 1998.

S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. C. Christiani, M. M. Wurfel, and X. Lin. 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, 2012.

L. Li, J. Ding, and G. R. Abecasis. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet*, 79:S2290, 2006.

W. Y. Lin, X. Y. Lou, G. Gao, and N. Liu. Rare variant association testing by adaptive combination of p-values. *PLoS One*, 9(1):e85728, 2014.

X. Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, 1997.

J. Liu, J. P. Lewinger, F. D. Gilliand, W. J. Gauderman, and D. V. Conti. Confounding and heterogeneity in genetic association studies with admixed populations. *Am. J. Epidemiol*, 177:351–360, 2013a.

L. Liu, D. Zhang, H. Liu, and C. Arendt. Robust methods for population stratification in genome wide association studies. *Bioinformatics*, 14(132), 2013b.

C. Loley, A. Ziegler, and I. R. Konig. Association tests for x-chromosomal markers–a comparison of different test statistics. *Hum Hered*, 71(1):23–36, 2011.

L. Luo, E. Boerwinkle, and M. Xiong. Association studies for next-generation sequencing. *Genome Research*, 21:1099 – 1108, 2011.

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

X. Mao, Y. Li, Y. Liu, L. Lange, and M. Li. Testing genetic association with rare variants in admixed populations. *Genet Epidemiol*, 37(1):38–47, 2013.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39:906–913, 2007.

E. R. Martin, S. A. Monks, L. L. Warren, and N. L. Kaplan. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet*, 67(1):146–154, 2000.

E. R. Martin, M. P. Bass, and N. L. Kaplan. Correcting for a potential bias in the pedigree disequilibrium test. *Am J Hum Genet*, 68(4):1065–1067, 2001.

I. Mathieson and G. McVean. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*, 44(3):243–246, 2012.

P. McCullagh and J. A. Nelder. *Generalized linear models*, volume 37. Chapman and Hall, New York, 1989.

C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437):pp. 162–170, 1997. ISSN 01621459.

S. A. Monks, N. L. Kaplan, and B. S. Weir. A comparative study of sibship tests of linkage and/or association. *Am. J. Hum. Genet*, 1998.

N. Mukhopadhyay, L. Almasy, M. Schroeder, W. P. Mulvihill, and D. E. Weeks. Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics (Oxford, England)*, 21(10):2556–2557, 2005.

B. M. Neale, M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly. Testing for an unusual distribution of rare variants. *PLoS Genet*, 7(3):e1001322, 2011.

J. Ott. Computer-simulation methods in human linkage analysis. *Proc Natl Acad Sci U S A*, 86(11):4175–4178, 1989.

K. Oualkacha, Z. Dastani, R. Li, P. E. Cingolani, T. D. Spector, C. J. Hammond, J. B. Richards, A. Ciampi, and C. M. Greenwood. Adjusted sequence kernel association test

for rare variants controlling for cryptic and family relatedness. *Genet Epidemiol*, 37(4): 366–76, 2013.

C. Papachristou, C. Ober, and M. Abney. Genetic variance components estimation for binary traits using multiple related individuals. *Genet Epidemiol*, 35(5):291–302, 2011.

J Pinheiro, D Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2013. R package version 3.1-113.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, 2006.

D. Qiao, M. Mattheisen, and C. Lange. On association analysis of rare variants under population substructure: An approach for the detection of subjects that can cause bias in the analysis - topt: An outlier detection method. *Genet Epidemiol*, 37(5):431–439, 2013.

D. Rabinowitz. A transmission disequilibrium test for quantitative trait loci. *Hum Hered*, 47(6):342–350, 1997.

D. Rabinowitz and N. Laird. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered*, 50:211–223, 2000.

A. A. Schaffer, M. Lemire, J. Ott, G. M. Lathrop, and D. E. Weeks. Coordinated conditional simulation with slink and sup of many markers linked or associated to a trait in large pedigrees. *Hum Hered*, 71(2):126–34, 2011.

S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*, 15(11):1576–83, 2005.

D. J. Schaid, S. K. McDonnell, J. P. Sinnwell, and S. N. Thibodeau. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet Epidemiol*, 37(5):409–18, 2013.

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

E. D. Schifano, M. P. Epstein, L. F. Bielak, M. A. Jhun, S. L. R. Kardia, P. A. Peyser, and X. Lin. Snp set association analysis for familial data. *Genet Epidemiol*, 36(8):797–810, 2012.

N. J. Schork, S. S. Murray, K. A. Frazer, and E. J. Topol. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*, 19(3):212–9, 2009.

K. D. Siegmund and B. Langholz. Ascertainment bias in family-based case-control studies. *American Journal of Epidemiology*, 155(9):875–880, 2002.

J. S. Sinsheimer, J. Blangero, and K. Lange. Gamete-competition models. *Am J Hum Genet*, 66(3):1168–1172, 2000.

J. S. Sinsheimer, C. A. McKenzie, B. Keavney, and K. Lange. Snps and snails and puppy dogs tails: analysis of snp haplotype data using the gamete competition model. *Ann Hum Genet*, 65(5):483–483, 2001.

S. L. Slager, D. J. Schaid, L. Wang, and S. N. Thibodeau. Candidate-gene association studies with pedigree data: controlling for environmental covariates. *Genet Epidemiol*, 24 (4):273–83, 2003.

D. J. Smith and A. J. Lusis. The allelic structure of common disease. *Hum Mol Genet*, 11 (20):2455–2461, 2002.

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

F. Sun, W. D. Flanders, Q. Yang, and M. J. Khoury. Transmission disequilibrium test (tdt) when only one parent is available: the 1-tdt. *Am J Epidemiol*, 150(1):97, 1999.

R. G. Svishcheva, N. M. Belonogova, and T. I. Axenovich. Ffbskat: Fast family-based sequence kernel association test. *PLoS ONE*, 9(6), 2014.

J. D. Terwilliger and J. Ott. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered*, 42(6):334–346, 1992.

T. Therneau, E. Atkinson, J. Sinnwell, D. Schaid, and S. McDonnell. *kinship2: Pedigree functions*, 2014. URL http://CRAN.R-project.org/package=kinship2. R package version 1.6.0.

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

T. Thornton and M. S. McPeek. Roadtrips: case-control association testing with partially or cocomplete unknown population and pedigree structure. *Am. J. Hum. Genet*, 86:172–184, 2010.

A. S. Turkmen and S. Lin. Blocking approach for identification of rare variants in family-based association studies. *PLoS One*, 9(1):e86126, 2014.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

V. Vieland and S. Hodge. Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. *Am J Hum Genet*, 56(1):33–43, 1995.

P. M. Visscher and D. L. Duffy. The value of relatives with phenotypes but missing genotypes in association studies for quantitative traits. *Genet Epidemiol*, 30:30–36, 2006.

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

Z. Wang and M. S. McPeek. An incomplete-data quasi-likelihood approach to haplotype-based genetic association studies onÂărelated individuals. *J Am Stat Assoc*, 104(487): 1251–1260, 2009.

R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439 – 447, 1974.

R. Wolfinger and M. O'connell. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:223–243, 1993.

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

Q. Yan, H. K.Tiwari, N Yi, G. Gao, K. Zhang, W. Lin, X. Lou, X. Cui, and N. Liu. A sequence kernel association test for dichotomous traits in family samples under a generalized linear mixed model. *Human Heredity*, 2015. (Unpublished).

D. Zhang and X. Lin. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1):57–74, 2003.

D. Zhang and X. Lin. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In DavidB. Dunson, editor, *Random Effect and Latent Variable Model Selection*, volume 192 of *Lecture Notes in Statistics*, pages 19–36. Springer New York, 2008.

Q. Zhang, L. Wang, D. Koboldt, I. B. Boreki, and M. A. Province. Adjusting family relatedness in data-driven burden test of rare variants. *Genet Epidemiol*, 38(8):722–727, 2014.

G. Zheng, J. Joo, C. Zhang, and N. L. Geller. Testing association for markers on the x chromosome. *Genetic Epidemiology*, 31(8):834–843, 2007.

X. Zhu, S. Li, R. S. Cooper, and R. C. Elston. A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet*, 82:352–365, 2008.

X. Zhu, T. Feng, Y. Li, Q. Lu, and R. C. Elston. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol*, 34(2):171–87, 2010.

Y. Zhu and M. Xiong. Family-based association studies for next-generation sequencing. *Am J Hum Genet*, 90(6):1028–45, 2012.