Statistical Methods for Detecting Rare Variant Associations in Family-Based Designs

By
Keng-Han Lin

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2017

Doctoral Committee:

Associate Professor Sebastian Zöllner, Chair
Professor Gonçalo Abecasis
Professor Michael Boehnke
Professor David Burke
Professor Trivellore Raghunathan

To those who have helped me along the way

To those who care about me and I care for

To my family who have been with me through everything

Acknowledgements

The journey toward my doctoral degree is extremely rewarding and grateful. It would not have been possible without the support and guidance from my committee, colleagues, friends and family.

First, I would like to express my deepest appreciation to my committee chair, Dr. Sebastian Zöllner, who introduces me the knowledge of statistical and population genetics. I especially enjoy the time discussing with him regarding various projects and he can always provide an insight and share his expertise to guide me on the right track. He is extremely patient, encouraging, and approachable. His excellent mentorship has fostered me as an independent scientist.

I would like to extend the gratitude to Dr. Michael Boehnke, who is always there to help and able to offer his advice for not only researching projects but also career guidance. I would also like to express my sincere thanks to Dr. Gonçalo Abecasis who has provided me valuable advices on my dissertation projects and the opportunity to participate in TOPMed project where I gain precious knowledge in large genetic studies. I would also like to thank Dr. Trivellore Raghunathan to generously offer his assistance to my projects, and thank Dr. David Burke for his encouragement to think outside the box and the help throughout my dissertation.

I would also like to thank the past and current members of the Zöllner lab: Yancy Lo, Rebecca Rothwell, and Jedidiah Carlson, for their generous exchange of scientific ideas,

friendships and time spent together. In addition, I would like to express my heartful appreciation to Kirsten Herold who has helped me becoming a better scientific writer.

Next, I would like to thank my friends, Hui-yu Yang, Sophie Chen, Yebin Tao, Linyao Jaden, and Adrian Tan. They have never stopped to show their support and care for me. My time in Ann Arbor have been filled with many joyful memories with these wonderful people.

Finally, I would like to thank my grandparents, Yu Lin and Xiu-Lan Su, my parents, Ming-Hung Lin and Shu-Ching Yang, and my brother, Keng-Li Lin, for their constant encouragement. My family has supplied me with their unbounded love; they are the foremost support to keep me motivated throughout the years in the United States.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

Rare variants are hypothesized to explain some genetic contribution to complex traits. However, using conventional case-control designs to identify rare variants associated with traits has low statistical power. Family designs can substantially increase power for these studies, especially for rare variants. In this dissertation, we present innovative statistical methods that can efficiently use family information to improve power.

In Chapter 2, we present TRAFIC, a rare-variant association test using affected sibpairs. For rare risk variants, two affected siblings would share the variant on their shared identity-by-descent (IBD) chromosomes. We thus test the distribution of rare variants on IBD chromosomes and non-IBD chromosomes. TRAFIC is robust to population stratification as “cases” and “controls” are matched within each sibpair. We show that TRAFIC has significant power gain over the population case-control design for variants with summed allele frequency < 5%. Considering allelic heterogeneity, where risk variants have different effect sizes, TRAFIC can double the power of a case-control study.

In Chapter 3, we present TRAP for testing the association between rare variants and a binary trait using extended families. Since affected family members are more likely to share risk variants, we propose to test if rare variants are shared more than expected given the known inheritance vector and the founder genotypes. TRAP is applicable from small to large pedigrees with multiple generations, including families with missing founders. Using simulations, we show that TRAP is more powerful than the conventional case-

control design and existing family-based approaches, especially for rare variants.

In Chapter 4, we present testing for rare variants associated with a continuous trait in extended families (TRACE). Given that a rare variant increases the trait value, conditional on inheritance vectors and founder genotypes, we test if family members with high trait values are more likely to share the variant. Under non-ascertained scenarios, TRACE can be more powerful than the existing family-based methods for large pedigrees; for ascertained scenarios, TRACE outperforms the existing approaches throughout considered pedigree structures. In sum, we present three family-based methods efficiently using the sharing of variants to increase the power for detecting rare variant associations.

Chapter 1
Introduction

Many human diseases are considered complex because they are attributable to a combination of genetic and environmental factors (Hunter 2005). While numerous studies have uncovered genetic and environmental factors associated with complex metabolic, cardiovascular and neurodegenerative diseases (McQueen et al. 2005; Burton et al. 2007; Halocarbon et al. 2007; Sladek et al. 2007; Hampe et al. 2007; McCarthy et al. 2008), the etiology of complex diseases have not been fully understood (Manolio et al. 2009; Eichler et al. 2010; Edwards et al. 2013). Moreover, these research findings have not been fully translated into effective clinical treatments (Wheeler et al. 2013; Ashley 2016; Nelson et al. 2016).

One characteristic of complex diseases is their tendency to be familial, meaning an individual often has an elevated risk of developing the disease in the presence of family history (McCarthy et al. 2008; Altshuler, Daly, and Lander 2008). Many diseases are measured in terms of their "heritability", which refers to the proportion of disease variance that is attributable to the genetic contribution (Manolio et al. 2009; Tenesa and Haley 2013). Inherited genetic variants are therefore presumed to play a significant role in the etiology of heritable diseases. Thus, the studies of complex diseases have been focusing on identifying genetic variants contributing to the disease heritability.

Linkage analysis was the first and predominant tool to map a genomic region (locus) containing variants contributing to the disease risk (Lander and Schork 1994; Altshuler, Daly, and Lander 2008). In a linkage study, researchers evaluate if there is a co-segregation of susceptible loci and the disease status in families. Using linkage studies, researchers successfully unraveled susceptible loci for many diseases with Mendelian inheritance pattern, where a single locus with large effect size determines the occurrence of diseases; examples including Huntington's disease and cystic fibrosis (Pericak-Vance 2001). Complex diseases, however, tend to have a genetic architecture inconsistent with the Mendelian inheritance (Risch and Merikangas 1996). The disease risk of a complex disease is hypothesized to be contributed by variants from many different loci, each having weak to modest effect. Therefore, when applied to complex disease studies, linkage analysis had low power and quickly lost its popularity (Lander and Schork 1994; Risch and Merikangas 1996; Altmüller et al. 2001; Ott, Wang, and Leal 2015). Furthermore, to assess the genetic contribution from multiple loci to complex diseases, a cost-effective genotyping technology was required to be applicable in a genome-wide scale.

With the advancement of genotyping array technology, which can assess thousands to millions of variants in large cohorts, genome-wide association studies (GWAS) quickly become the dominant choice for researchers to study the genetic contribution to complex diseases. In a GWAS, a sample of cases (i.e. people with the disease of interest) and controls (people without the disease) from a population are genotyped and statistically tested to determine if any variants are more frequently observed in cases than controls. GWASs have successfully revealed numerous genetic variants attributable to many complex diseases and traits, such as diabetes, psychiatric disorders, cancers, height, and

blood pressure. By the year 2015, > 2,000 GWAS publications have identified > 15,000 variants associated with > 600 human diseases and traits (Welter et al. 2014). Although a large number of associated variants have been identified, these variants only collectively explain a modest proportion of heritability for a given disease (Manolio et al. 2009); this discrepancy between the observed and the estimated heritability is termed as missing heritability. One of the explanations for the missing heritability is that the variants identified in GWAS are mostly common variants with frequency > 5% in the population (Manolio et al. 2009). It is believed that a risk variant reducing the fitness (causing a deadly disease) is less likely to be transmitted to offspring, resulting in a low frequency in the population; therefore, it is hypothesized that rare variants undetected by array genotyping could explain a substantial portion of the missing heritability (Manolio et al. 2009; Eichler et al. 2010; Cirulli and Goldstein 2010; Lee et al. 2014). However, the discovery of rare variants requires much denser genotyping, which was not feasible with array technologies.

it was not until the innovation of next-generation sequencing technology in 2005 that the comprehensive discovery of rare variants has become feasible (Goodwin, McPherson, and McCombie 2016). Using next-generation sequencing, a massive number of rare variants have been identified. In the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015), which sequenced the full genome of > 2,500 individuals from world-wide populations, the majority of variants were rare: 72% of 88 million variants have allele frequency < 0.5%. In addition, sequencing protein-coding regions (~1% of the human genome) of 60,706 individuals of diverse ancestries has identified 99% of 10 million variants with allele frequency < 1%, and 54% are singleton variants (only observed in one

individual; allele frequency ~0.002%) (Lek et al. 2016). Identifying the association between these rare variants and diseases typically requires a large number of samples; > 10,000 samples would be necessary to statistically detect an associated variant with frequency 0.1% and modest effect size (Manolio et al. 2009). In addition, the sample size is roughly quadratic to the inverse of effect size, suggesting finding an extremely rare variant association with weak effect size will require an enormous number of samples. Although the cost of sequencing has dropped to ~$1,000 per sample in 2015 (Goodwin, McPherson, and McCombie 2016), it still poses a substantial financial burden to collect such a large sample size. Furthermore, with a large sample size, the heterogeneity in samples can easily confound results and generate spurious signals, known as population stratification, in which samples from different populations consist of different allele frequency and disease prevalence (Zawistowski et al. 2014). Methods that controls for population stratification for common variants may also not reliably reduce the confounding effect caused by population stratification for rare variants (Mathieson and McVean 2012). Thus, it is important to develop new tools and efficient study designs to facilitate the process of untangling the rare variant contribution toward complex diseases.

One possible solution to adjust for population stratification is to use family samples where family members typically have a matched genetic and environmental background (Ott, Kamatani, and Lathrop 2011). When using only within-family information, a family-based association test can be robust to population stratification (Spielman, McGinnis, and Ewens 1993; Abecasis, Cardon, and Cookson 2000). Another advantage of using family samples is to increase the observed copies of risk variants in the susceptible loci, leading to power gain (Ott, Kamatani, and Lathrop 2011). For example, for a rare risk variant,

multiple affected family members in a family are likely to carry the same variant. Because they can be powerful and robust to population stratification, family-based study designs are an attractive and tractable means of studying the relationship between rare variants and complex diseases.

Recently proposed family-based methods for rare-variant associations typically extend the existing methods to gene-based tests, which combine the information across rare variants in a gene region to collectively assess their associations (H. Chen, Meigs, and Dupuis 2013; Ionita-Laza et al. 2013; Schaid et al. 2013; De et al. 2013); these methods can be categorized into two classes: The first class separate extended families into many trios as proposed by FBAT (Laird, Horvath, and Xu 2000) and adapts the gene-based tests (Ionita-Laza et al. 2013; De et al. 2013). This class of approaches loses efficiency from not jointly examining an extended family and not considering founders' disease status (Laird and Lange 2006). The second class of approaches directly extends population gene-based tests to adapt family samples by accounting for the relatedness (Schifano et al. 2012; H. Chen, Meigs, and Dupuis 2013; X. Wang et al. 2013). However, the adjustment to the relatedness reduces the effective sample size, therefore not efficiently using the information contained in the family, offsetting the benefit of enriching the genetic loading at the susceptible loci (Lin and Zöllner 2015). Furthermore, existing family-based methods are typically applicable to specific pedigree structures or can only apply to either binary or continuous traits. In this dissertation, we present efficient family-based methods for rare variant associations that can apply to both binary and continuous traits, and be flexible to arbitrary pedigree structures, from sibling pairs (sibpairs) to extended pedigrees.

In Chapter 2, we present a new method for association testing based on affected sibpairs. For rare risk variants, an affected sibpair is more likely to share the same risk variant identity-by-descent (IBD), i.e. the variant resides on the shared IBD chromosomes. Based on this observation, we test for the association in affected sibpairs by comparing the allele count on shared IBD chromosome to non-shared IBD chromosome, and develop a statistical test, which we refer to as test for rare-variant association with family-based internal control (TRAFIC) (Lin and Zöllner 2015). TRAFIC is generally robust to population stratification as "cases" and "controls" are matched within each sibpair. By using simulations, we show that TRAFIC has the most significant power gain over the population case-control design for rare variants with summed allele frequency < 5% in a given genomic region. Considering allelic heterogeneity, where risk variants have different effect sizes, TRAFIC doubles the power of a case-control study in many realistic parameter settings.

Many recent family-based studies collect data from extended families (Hunt et al. 2005; Mahmood et al. 2014; Sidore et al. 2015). With common diseases, including unaffected family members can provide additional information (Laird, Horvath, and Xu 2000). Using the full information from extended families can further improve power (Laird and Lange 2006). However, few methods have used IBD information for extended families. Sul et al. [2016] proposed RareIBD to evaluate the excessive sharing of IBD variants among affected and the depletion of sharing on unaffected family members. To reduce the computational burden while evaluating the significance, RareIBD assumes that only one founder in each family pedigree carries the risk variant. However, when multiple family members are ascertained to be affected, this assumption may not be valid. Therefore,

RareIBD has inflated type I error in these ascertained families (Sul et al. 2016). Thus, devising a new and efficient family-based methods that can correctly harness extended families is needed.

Therefore, in Chapter 3, we present testing for rare-variant associations using pedigree studies (TRAP). A variant that increases disease risk is more likely to be transmitted from founders to multiple affected offspring and resides on IBD chromosomes. Thus, conditional on the inheritance vector, TRAP evaluates the departure from the expected number of variants shared among affected and among unaffected family members. TRAP is applicable to any type of pedigree structures, from sibpairs to multi-generation families, and can also apply to families with missing founders. We evaluate TRAP using simulations with realistic parameter configurations and show that, given the same number of sequenced individuals, TRAP can substantially outperform the conventional population case-control design and existing family-based methods.

Both TRAFIC and TRAP described above are applicable to test for rare variant association with binary traits. However, continuous traits are routinely and readily collected in many studies; for example, in a diabetes study, patients are often measured for their lipid level, glucose level, waist/height ratio, and blood pressure. Existing family-based methods for studying continuous traits typically face the same challenge as the methods for studying binary traits, namely not efficiently using the family information. In addition, different sampling strategies can affect the efficiency of these existing family-based methods for continuous traits. There are two common ways of sampling families: the first design is to collect families ascertained to have extreme traits; if a rare variant increases the trait value, this rare variant is likely to segregate in family members who

have a high-level trait. Collecting such ascertained families can increase the observed frequency of associated variants, leading to power gain. Although powerful, the ascertainment requirement typically increases the cost of collecting such families, resulting in a smaller sample size (given a fixed budget), which can offset the increased copies of associated variants. The second design is to collect a larger sample of unascertained families, i.e. randomly selected families in a population. This design may still allow us to observe excessive copies of risk variants within a family, described as the 'Jackpot' effect (Feng et al. 2015). Thus, developing a method that provides adequate statistical performance for both scenarios can ease the process of identifying rare variant associations and not have to consider whether the family is "ascertained" or not.

In Chapter 4, we present an efficient method to test for rare variant associations with continuous traits using extended families (TRACE), which can be applied to both non-ascertained and ascertained scenarios. TRACE is a combined test that use both within-family and between-family information. For within-family information, we extend the idea from TRAP that if a variant increases the trait value, we test if family members with high trait values are more likely to share the variant. Equivalently, conditional on the inheritance vector, the variant would pass through the founder chromosome to include many family members with high trait values. To exploit between-family information, we test for the association between the genetic loading and the mean trait of a family. Using simulations, we show that including between-family information can substantially improve power compared to only using within-family information. Under a non-ascertained scenario, TRACE is more powerful than the existing family-based methods for large pedigree; for ascertained scenarios, TRACE shows an advantage in power over the

existing approaches in all considered pedigree structures.

In summary, we develop three innovative methods in this dissertation to efficiently test for rare variant associations with complex diseases by using family samples. Our methods provide practical sampling guidelines for future large-scaled family studies which can achieve the highest power to detect associations. Identifying associated rare variants to complex diseases can facilitate the downstream steps to study their functionality, thereby developing advanced strategies for early detection and treatments.

Chapter 2
Robust and Powerful Affected Sibpair Test for Rare Variant Associations

2.1. Introduction

Rare variants with large relative risk are hypothesized to explain some of the missing heritability of complex diseases (Manolio et al. 2009). Several studies have identified rare variants underlying rare Mendelian diseases using next-generation sequencing technology (Ng et al. 2009; Ng et al. 2010). However, the conventional case-control design has low statistical power to detect the association between rare variants and complex diseases (B. Li and Leal 2008; Cooper and Shendure 2011). To overcome the low power of single-marker test on rare variants, researchers have proposed to combine variants in a gene or genomic region to test for association (B. Li and Leal 2008; Zawistowski et al. 2010; Price et al. 2010). However, such gene-based tests in population samples may still need >10,000 individuals to identify the signal from rare variants (Nelson et al. 2012); sequencing such large samples is still very expensive. Moreover, large samples are typically more heterogeneous in origin, increasing the risk of population stratification (Price et al. 2006). In such large samples, even subtle stratification causes substantially

increased false positive rate in rare variant tests (Zawistowski et al. 2014). While methods to control for population stratification, such as principal components and genomic control (Devlin and Roeder 1999; Price et al. 2006) have been successfully applied to common variants, it is unclear whether such methods are appropriate for rare variant tests (Mathieson and McVean 2012; Liu, Nicolae, and Chen 2013).

As family members are naturally matched for genetic background, several recent gene-based methods for testing the association between rare variants and the phenotype adapt family data to control for population stratification (Guo and Shugart 2012; De et al. 2013) . In addition, the allele frequency of rare risk variants in cases can be substantially increased by collecting cases with affected relatives (Fingerlin, Boehnke, and Abecasis 2004; Peng et al. 2010; Zöllner 2012). While collecting families with multiple affected members is challenging, family-based studies of rare variants can leverage existing large collections of families that were originally generated for linkage analysis (Rao et al. 2003; Howson et al. 2009; Guan et al. 2012); for example, International Type 2 Diabetes Linkage Analysis Consortium contains > 4000 affected sibpairs (Guan et al. 2012).

Methods have been proposed to extend the current collapsing tests to rare variants in family data. Guo and Shugart (2012) and De et al. (2013), extended the family-based association test (FBAT) (Laird, Horvath, and Xu 2000) to rare variants in the style of a collapsing test. Schifano et al. (2012) and Chen et al. (2013) used linear mixed models to extend the SNP-set kernel association test (SKAT) (Wu et al. 2011) to families. Shugart et al. (2012) and Fang et al. (2012) proposed to estimate the relatedness between samples and adjust the test statistics for rare variant association accordingly. However, none of the existing methods directly leverage the benefit of studying families where the same rare

variant is observed multiple times. By using such information, we can increase power to detect the association between rare variants and the phenotype.

Here, we propose a powerful framework for testing rare variant associations using affected sibpairs. We create a matched design by comparing the allele count of rare variants on shared identical by descent (IBD) chromosome regions to the allele count on non-shared identity by descent chromosome regions across affected sibpairs in a region of interest. Sharing status of chromosome regions can be easily estimated using high density genotype data (Keith et al. 2008), and sharing status of alleles can be inferred conditional on the known chromosome region sharing status. Intuitively, we consider shared chromosome regions as "case" chromosome regions and non-shared chromosome regions as "control" chromosome regions. Under the null hypothesis of no association, the probability of a shared chromosome region carrying an allele is identical to the probability of a non-shared chromosome region carrying an allele. Under the alternative that an allele increases/decreases the disease risk, the probability of a shared chromosome region carrying that allele is higher/lower than the probability of a non-shared chromosome region carrying that allele.

We evaluate this design by calculating the analytical power for a collapsing gene-based test (B. Li and Leal 2008), assuming a general model of rare risk alleles that is specified by the summed allele frequency of all rare risk variants in the gene and the mean and variance of their effect size (Zöllner 2012). We show that given the same number of sequenced individuals, the power of the proposed affected sibpair test for rare-variant association with family-based internal control (TRAFIC) is higher than the conventional case-control design for rare risk variants (summed risk allele frequency < 0.05).

Considering allelic heterogeneity, where risk variants have different effect sizes, TRAFIC doubles the power of a case-control study in many realistic parameter values. We also evaluate the power of the proposed method under various gene-gene interaction models and find that power depends on the type of interaction and the overall heritability of the disease. Using simulations, we also show that the proposed TRAFIC is generally robust to population stratification.

2.2. Materials and Methods

2.2.1. Test for rare-variant association with family-based internal control (TRAFIC)

We consider a set of affected sibpairs with known number of chromosome regions shared identical by descent (IBD). At a locus of interest (for example a gene), we compare the count of alleles of rare variants on chromosome regions shared IBD between the siblings to the count of alleles of rare variants on chromosome regions not shared IBD (non-IBD chromosome regions) across sibpairs. Let, p^{IBD} be the frequency of IBD chromosome region carrying at least one allele and p^{NonIBD} be the frequency of non-IBD chromosome regions carrying at least one allele. Alleles without effect on disease risk are equally likely to occur on any chromosome region regardless of IBD status. Thus, the null hypothesis under no association is $H_0: p^{IBD} = p^{NonIBD}$. Variants that are associated with the phenotype (protective or causative) would differ in frequency between IBD and non-IBD chromosome regions. Hence, we can test for departure from the null hypothesis either in a collapsing framework by considering the alternative $H_a: p^{IBD} \neq p^{NonIBD}$ or in a dispersion framework where this alternative is considered for each variant and the

combined test statistic aggregates the evidence across all variants.

	0 IBD chromosome region	1 IBD chromosome region	2 IBD chromosome regions
Both siblings are homozygous minor allele	4 non-shared alleles	1 shared and 2 non-shared alleles	2 shared alleles
One homozygous minor allele and one heterozygote	3 non-shared alleles	1 shared and 1 non-shared alleles	N/A
Both siblings are heterozygous	2 non-shared alleles	Ambiguous configuration	1 shared allele

Table 2-1. Identification of variant IBD status conditional on chromosome region IBD status. Assuming chromosome region IBD status is known, the number of shared and non-shared alleles can be inferred for all but one configuration of genotypes (shaded cell).

In a sibpair with known IBD status, identifying whether an allele of a variant is located on an IBD or a non-IBD chromosome region is straightforward for most genotypes as shown in Table 2-1; for example, when a sibpair does not share the chromosome region (0 IBD chromosome region), all observed alleles for that variant in two siblings are non-shared; for a sibpair who shares 1 IBD chromosome region, the alleles of a homozygous sibling must be one shared and one non-shared. Only when the sibpair shares one IBD chromosome region and the genotypes are heterozygous in both individuals, the IBD status of the allele is ambiguous (shaded in Table 2-1): this configuration could be either the result of a single rare allele located on the IBD chromosome region or two copies of the rare allele inherited separately on the non-IBD chromosome regions (as illustrated in Figure A-1).

To resolve this ambiguous configuration, we implement an imputation algorithm and use simulations to show the false positive rate is controlled (see Appendix A.1 for details).

2.2.2. Evaluating TRAFIC

The analytical power of the proposed TRAFIC based on a collapsing gene-based test depends on the difference between the expected allele count on shared IBD chromosome regions and the expected allele count on non-shared IBD chromosome regions. To calculate these expectations, we assume that all rare variants evaluated in a locus occur on different haplotypes. Let f be the sum of population allele frequencies of all risk variants (summed risk allele frequency). For each sibpair, we count the number of alleles $\mathrm{H}_S \in \{0,1,2\}$ on the shared chromosome regions and the number of alleles $\mathrm{H}_{NS} \in \{0,1,2,3,4\}$ on non-shared chromosome regions. Let AA_R be an affected sibpair and $\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}|\mathrm{AA}_R, \mathrm{S})$ be the probability of $\mathrm{H}_S, \mathrm{H}_{NS}$ conditional on the number of shared IBD chromosome regions $\mathrm{S} \in \{0,1,2\}$. Using Bayes' rule, we can write this conditional probability as

$$\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}|\mathrm{AA}_R, \mathrm{S}) = \mathrm{P}(\mathrm{AA}_\mathrm{R}|\mathrm{H}_S, \mathrm{H}_{NS})\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}|\mathrm{S})\mathrm{P}(\mathrm{S})\frac{1}{\mathrm{P}(\mathrm{AA}_R, \mathrm{S})},$$

where $\mathrm{P}(\mathrm{AA}_R|\mathrm{H}_S, \mathrm{H}_{NS})$ depends on the underlying genetic and effect size model (see Appendix A.3 for derivations). Based on previous work (Zöllner 2012), we model the effect size (relative risk) of each risk haplotype as a random variable with the first two moments μ and σ^2. Then, $\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}|\mathrm{AA}_R, \mathrm{S})$ is fully determined by the parameters μ, σ^2, and f. We calculate the power for TRAFIC based on $\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}|\mathrm{AA}_R, \mathrm{S})$ for a range of relative risk parameter μ and σ^2, and under different f assuming a simple collapsing method (B. Li and Leal 2008) to test the association between rare variants and the dichotomous phenotype (more details in Appendix A.4). To maintain an overall false positive rate of 0.05 after testing 20,000 genes in the genome, we set the false positive rate to 2.5×10^{-6}. We compare our proposed TRAFIC with two other designs: (1) the conventional case-control study

comparing a sample of cases to unaffected controls. (2) A selected cases design comparing cases that are ascertained to have an affected sibling to unaffected controls (Fingerlin, Boehnke, and Abecasis 2004; Zöllner 2012). All designs retain the nominal false positive rate under the null (Table A-1).

2.2.3. Simulation setup for TRAFIC

To validate the derived analytical results, we simulate sibpair samples and apply our proposed TRAFIC. We first generate four independent parental haplotypes, each carrying a risk allele with probability *f*. Without considering recombination, we then generate two descendants, each randomly inheriting one chromosome region from each parent. Following Risch (1990), we define the contribution to prevalence *K* at the locus of interest as K_L and the contribution of the remaining genome as K_G. The prevalence among subjects with an affected relative with relation status R is K_R; the contribution to K_R at the locus of interest and the remaining genome are then K_{LR} and K_{GR}, respectively. We adjust $K_G K_{GR}$ under the multiplicative model to maintain both K and the sibling relative risk (SRR).

$$SRR = \frac{K_L K_{LR} K_G K_{GR}}{K \times K}.$$

Here $K_L K_{LR}$ depends on $\mathrm{P}(\mathrm{AA}_R | \mathrm{H}_S, \mathrm{H}_{NS})$ (more details in Appendix A.3). The relative risk of the risk allele follows a gamma distribution with specified μ and σ^2. Thus, the probability of having both siblings in the family affected is $K_L K_{LR} K_G K_{GR}$ and is set to 1 if the simulated probability exceeds 1. We generate datasets of 1000 affected sibpairs in each replicate. To evaluate the performance of our multiple imputation algorithm, we generate sibpairs assuming the sharing status is known. Then we mask the true location

for the double-heterozygote sibpairs who share one IBD chromosome region and apply our multiple imputation algorithm.

2.2.4. Population stratification

Using the simulation design described above, we evaluate the impact of population stratification. We simulate two populations with summed risk allele frequency of 0.01 and 0.05, respectively, and assign a prevalence ratio π between two populations. Assuming two populations have the same sibling relative risk, the ratio of frequencies of affected sibpairs between the two populations is then π^2. Assuming that both populations contribute equally, we generate case-control samples by sampling 1000 cases, a proportion of $\pi/(1+\pi)$ from population 1 and $1/(1+\pi)$ from population 2. We also sample 1000 controls with equal contribution from each population. To generate a stratified sample for TRAFIC, we generate a sample of 1000 affected sibpairs with a proportion of $\pi^2/(1+\pi^2)$ from population 1 and a proportion of $1/(1+\pi^2)$ from population 2. We assume unknown sharing status for double-heterozygote sibpairs who share one IBD chromosome region and impute the sharing status through multiple imputation. To generate cases for the selected cases design, we sample affected sibpairs with a proportion of $\pi^2/(1+\pi^2)$ from population 1 and $1/(1+\pi^2)$ from population 2; controls are sampled evenly from both populations. We generate 1000 datasets for each value of π and estimate the false positive rate.

2.2.5. Gene-gene interaction

Interaction between the locus of interest and the remaining genome can influence the power of association tests in family samples (Zöllner 2012). We model gene-genome

interaction as two loci, L and G. L is the locus of interest while G represents genetic effects in the remainder of the genome. We define the joint effect as

$$P(A|h_m, h_n, g_s, g_t) \propto \beta_L^{h_m+h_n} \beta_G^{g_s+g_t} \gamma^{(h_m+h_n)(g_s+g_t)}.$$

where h_m and h_n represent the indicator of a risk allele at locus L; let g_s and g_t represent the indicator of a risk allele at locus G. In the absence of risk alleles at G, all risk alleles at locus L have the same relative risk β_L . Moreover, we describe the extent of interaction in this model by the parameter γ as the relative risk when risk alleles are present at both loci L and G, where $\gamma = 1$ indicates no interaction, $\gamma < 1$ indicates antagonistic interaction, and $\gamma > 1$ indicates synergistic interaction. Under this model, the marginal relative risk at locus L is

$$\frac{P(A|h_m = 1)}{P(A|h_m = 0)} = \frac{\beta_L \sum_{h_n} \beta_L^{h_n} p(h_n) \sum_{g_s} \sum_{g_t} (\beta_G \gamma)^{g_s + g_t} \gamma^{h_n(g_s+g_t)} p(g_s) p(g_t)}{\sum_{h_n} \beta_L^{h_n} p(h_n) \sum_{g_s} \sum_{g_t} \beta_G{}^{g_s + g_t} \gamma^{h_n(g_s+g_t)} p(g_s) p(g_t)}.$$

The marginal relative risk at locus G is expressed in a similar fashion. To explore the effect of gene-gene interaction, given the sibling relative risk, we vary γ while adjusting β_L and β_G to keep the marginal relative risks constant (see Appendix A.5). This maintains a constant power for the conventional case-control study. We then calculate $\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}|\mathrm{AA}_\mathrm{R})$ at locus L and evaluate the power of TRAFIC for different values of γ.

2.2.6. An example to illustrate TRAFIC

To illustrate how to apply TRAFIC, we simulate 1000 sibpairs assuming the number of shared IBD chromosome region is known. We simulate sequence data by using coalescent-model based simulator COSI (Schaffner et al. 2005) to generate a population of ten thousand 1kb haplotypes. From the 50 variants in the region, we randomly pick 10 variants with minor allele frequency (MAF) < 0.05 and assign each variant the relative risk

as a function of MAF, -$\log_{10}$(MAF) (Wu et al. 2011). In this setting, a variant with MAF = 0.05 has relative risk of 1.33 and a singleton has relative risk of 4. We thus generate a population with $f = 0.025$, $\mu = 2.52$, and $\sigma^2 = 0.62$; then, we generate 1000 affected sibpairs and apply TRAFIC to that dataset.

The simulated data contains 254, 509 and 237 sibpairs who share 0, 1, and 2 chromosome regions, respectively; these equal to 983 shared chromosome regions and 2034 non-shared chromosome regions. Excluding 42 sibpairs who shared one chromosome region with ambiguous double-heterozygote genotypes, there are 51 shared and 67 non-shared chromosome regions carrying at least one allele (carrier). Using imputation to resolve the IBD status of allele from 42 sibpairs with ambiguous double-heterozygote genotypes, the mean count of carrier chromosome regions is 91.7 on shared chromosome regions and 67.6 on non-shared chromosome regions. Using a χ^2 test, we reject the null hypothesis that IBD and non-IBD chromosome regions are equally likely to carry at least one allele ($\mathrm{p} = 5.63\times10^{-11}$) indicating the presence of risk variants at this locus.

2.3. Results

We proposed a new gene-based method for analyzing affected sibpairs by comparing the risk alleles on shared IBD chromosome regions with the risk alleles on non-shared IBD chromosome regions. We evaluated the proposed TRAFIC design assuming a collapsing gene-based test by modeling allelic heterogeneity at the locus of interest based on a summed allele frequency of all risk variants *f* and a distribution of effect sizes with mean μ and variance σ^2. For comparison, we also evaluated the conventional cases-control

design (conventional) and a case-control design in which the cases are selected conditional on having an affected sibling (selected cases) under the same genetic model. For all three designs, we assumed equal number of sequenced or genotyped individuals. To use consistent language, we referred to shared IBD chromosome regions in TRAFIC as cases and non-shared IBD chromosome regions as controls.

First, we compared the expected summed minor allele frequency (sMAF) in cases and controls with and without allelic heterogeneity to illustrate how TRAFIC behaved relative to the conventional and selected cases designs. We then calculated the analytical power of three designs. We also evaluated robustness to population stratification. Finally, we calculated the analytical power of TRAFIC while considering different directions of gene-gene interaction.

2.3.1. Frequency distribution of risk variants

To quantify the enrichment of risk variants in TRAFIC, we calculated the expected summed minor allele frequency (sMAF) of risk variants in cases and controls of TRAFIC for a range of genetic models (see Appendix A.4 for details). Initially, we modeled a locus with constant genetic risk μ between 1 and 5 for all variants ($\sigma^2 = 0$) (Figure 2-1) and a disease prevalence of 0.01. In TRAFIC (Figure 2-1a), sMAF increased rapidly in cases (shared IBD chromosome regions) and also increased roughly linearly with μ in controls (non-shared IBD chromosome regions). In the conventional design (Figure 2-1b), sMAF increased in cases almost linearly with relative risk, only slightly faster than the sMAF in controls of TRAFIC. In the selected cases design (Figure 2-1c), sMAF in cases with affected siblings increased faster than cases in the conventional case-control design but slower than sMAF in cases of TRAFIC. Both in the conventional design and the selected cases

design, sMAF in controls decreased slightly as μ increased, especially for more common variants ($f = 0.20$). As a result, TRAFIC generated a larger difference in sMAF between cases and controls than the conventional case-control design in models with $f = 0.01$ and 0.05. This advantage of TRAFIC reduced with increasing *f*. For μ = 2, the difference in sMAF of TRAFIC compared to the conventional design was 190% (0.019 to 0.010) at *f* = 0.01 and reduced to 123% (0.166 to 0.135) at *f* = 0.20. For a higher disease prevalence of 0.20, the sMAF in controls decreased more rapidly as μ increased and the difference between cases and controls grew further in the conventional case-control and selected cases design (Figure A-2).

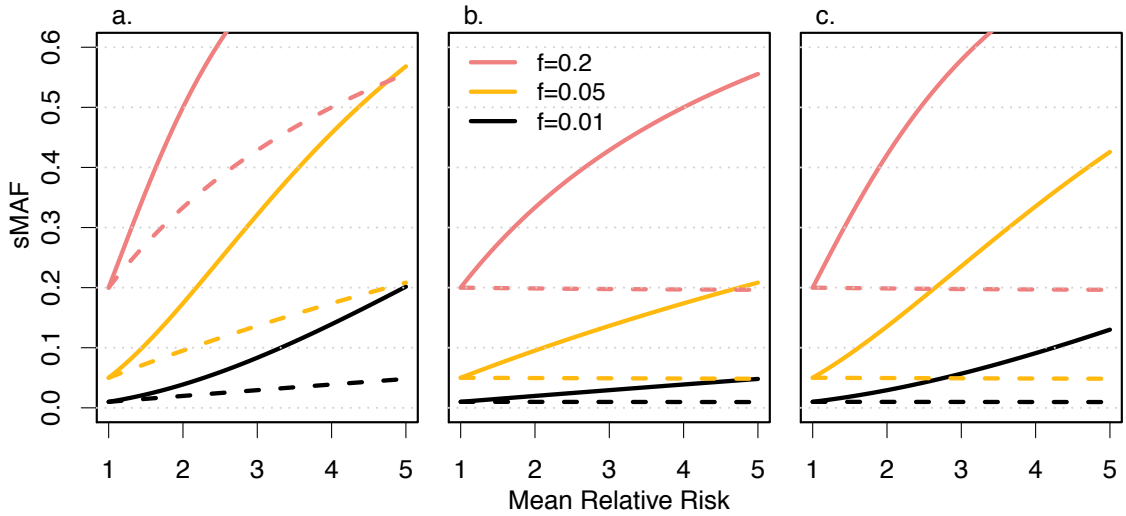


Figure 2-1. Summed minor allele frequency (sMAF) of risk variants in cases (solid lines) and controls (broken lines) under different study designs. We show sMAF as a function of mean relative risk of risk variants for (a) TRAFIC, (b) the conventional case-control design, and (c) the selected cases design for summed allele frequencies (f) of 0.01, 0.05 and 0.2 and fixed variance of relative risk $\sigma^2 = 0$.

To evaluate scenarios where genetic effect differs between risk variants, we considered a distribution of relative risks with $\sigma^2 > 0$ while maintaining $\mu = 1.5$ (Figure 2-2); for $f = 0.01$, a value $\sigma^2 = 0.1$ represents e.g. a scenario of 20 tested variants with equal frequencies where 6 of the tested variants are non-functional (relative risk = 1) and 14 of the tested variants have a relative risk of 1.71. A value $\sigma^2 = 0.2$ would e.g. represent

9 non-functional variants and 11 variants with relative risk 1.91. Increasing σ^2 did not affect sMAF in cases or controls in the conventional design, as in this design sMAFs only depended on μ (Figure 2-2b). In TRAFIC, sMAF in cases increased with σ^2 while the sMAF in controls remained constant. Similarly, in the selected cases design, sMAF in cases increased with σ^2, albeit more slowly than for TRAFIC (Figure 2-2a and c). Even if the average effect of risk variants is 1 ($\mu = 1$), the difference in sMAF between cases and control increased with growing σ^2 for TRAFIC and for the selected cases design (Figure A-3).

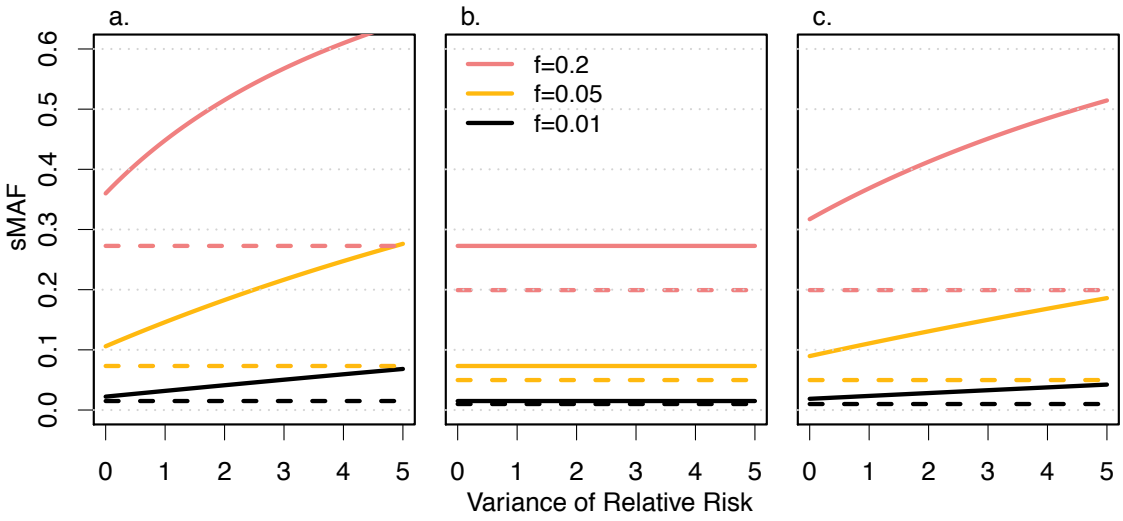


Figure 2-2. Summed minor allele frequency (sMAF) of risk variants in cases (solid lines) and controls (broken lines) under different study designs. We show sMAF as a function of variance of relative risk between risk variants for (a) TRAFIC, (b) the conventional case-control design, and (c) the selected cases design for summed allele frequencies (f) of 0.01, 0.05 and 0.2 and fixed mean relative risk $\mu = 1.5$.

2.3.2. Power Analysis

Based on the differences in expected sMAF, we calculated the analytical power for three study designs for the same number of individuals (n=2000): (1) 1000 affected sibpairs using TRAFIC, (2) 1000 cases and 1000 controls in the conventional cases-control design, and (3) 1000 cases with affected siblings and 1000 controls in the selected cases design. Thus, we generated 4000 independent observations for the conventional and the

selected design, and ~3000 independent observations (~1000 cases and ~2000 controls) for TRAFIC. We also determined power empirically using simulations and observed no difference between empirical power and analytical power (Figure A-4).

Assuming all risk variants had the same relative risk between 1 and 5 ($\sigma^2 = 0$), the selected cases design was uniformly most powerful (Figure 2-3a) while the power ranking of TRAFIC and the conventional design depended on *f*. For rarer risk variants ($f < 0.05$), TRAFIC had substantially higher power than the conventional design across all relative risks analyzed. For example, for $f = 0.01$ and $\mu = 2.5$, the power of the conventional design was 0.131 compared to 0.532 for TRAFIC. With increasing f or increasing prevalence, the power difference between TRAFIC and the conventional design reduced. For sets of risk variants with $f > 0.05$, the power of the conventional design was larger than the power of TRAFIC. The ranking of TRAFIC with the conventional design depended on the prevalence of the disease, for prevalence 0.20, the conventional design was already more powerful than TRAFIC for $f > 0.01$ (Figure A-5).

In models with allelic heterogeneity ($\sigma^2 > 0$), power of TRAFIC increased with rising σ^2 while the power of the conventional design was independent of σ^2 and only depended on f (Figure 2-3b). For $f = 0.01 \text{ and } 0.05$ at $\mu = 1.5$, the power of TRAFIC was uniformly greater than the power of the conventional design. For $f = 0.2$, TRAFIC was more powerful than the conventional design for $\sigma^2 > 0.1$. Even for high-prevalence diseases, TRAFIC was more powerful than the conventional design at modest levels of heterogeneity (Figure A-5). Moreover, the selected cases design was no longer uniformly most powerful in the presence of moderate allelic heterogeneity. For example, when $f = 0.01$ and $\sigma^2 = 2$, TRAFIC outperformed the selected cases design (with power of 0.412

and 0.306, respectively). For a model with no mean effect ($\mu = 1$), TRAFIC was uniformly the most powerful regardless of f (results not shown).

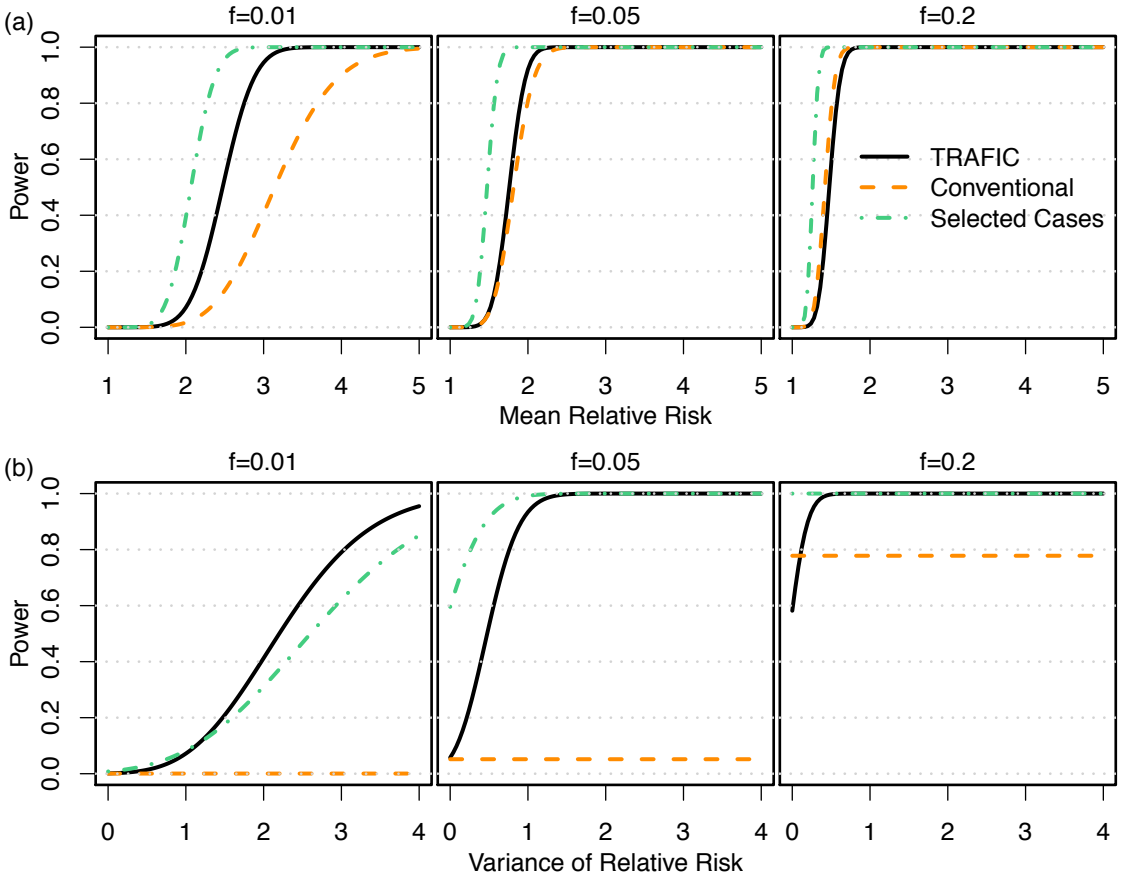


Figure 2-3. The analytical power curve for TRAFIC, conventional case-control, and selected cases design for different summed allele frequencies (f).Row (a) displays the power as a function of mean relative risk evaluated at variance of relative risk $\sigma^2 = 0$. Row (b) shows the power as a function of variance of relative risk evaluated at mean relative risk $\mu = 1.5$. Results are shown for 2000 individuals (1000 sibpairs or 1000 cases and 1000 controls) at a significance level 2.5×10^{-6}.

2.3.3. Population stratification

We modeled the level of population stratification by the parameter π which represents the ratio of prevalence between two populations (see methods). In the absence of true risk variants ($\mu = 1, \sigma^2 = 0$), the conventional case-control design and the selected cases design only achieved the nominal false positive rate at $\pi = 1$ where equal proportion of cases and controls were sampled from the two populations. Both

designs showed substantially increased false positive rate when moving away from $\pi = 1$. Especially the selected cases design showed a high false positive rate for moderate levels of stratification. For $\pi = 1.22$, the false positive rate was 0.064 and 0.107 for the conventional case-control and selected cases designs; the inflation increased to 0.725 and 0.973 when $\pi = 4.06$. TRAFIC maintained the false positive rate at the nominal level of 0.05 across the range of π (Figure 2-4) as long as we assumed either no linkage signal or a linkage signal of the same strength in the two populations. When we modeled a strong linkage signal in only one of the populations, we observed a slightly increased false positive rate in TRAFIC (Appendix A.6).

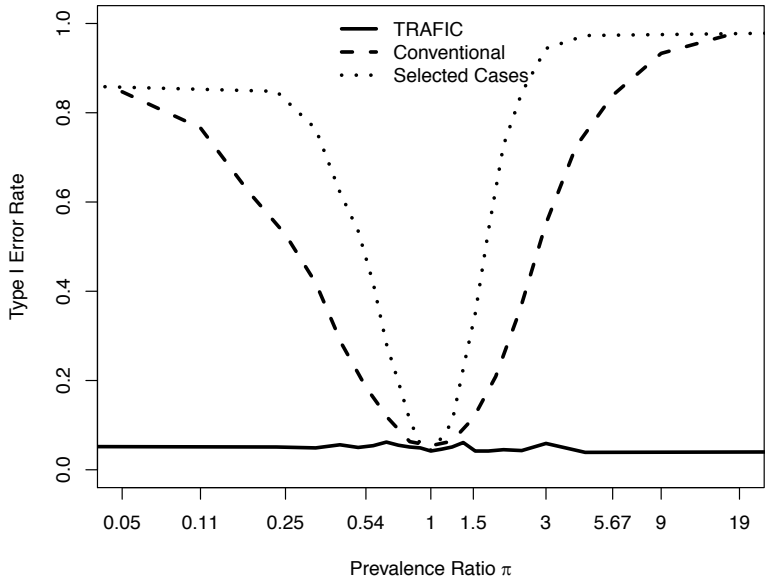


Figure 2-4. False positive rate in the presence for population stratification for TRAFIC, selected cases and the conventional case-control design. The false positive rate is shown as a function of the prevalence ratio π between two sampled populations. Calculations are based on a summed allele frequency of 0.01 in population 1 and 0.05 in population 2, and a sample size of 2000 individuals (1000 sibpairs or 1000 cases and 1000 controls) at a significance level 0.05.

2.3.4. Gene-gene interaction

We summarized the effect of the gene-gene interaction in a two-locus model by the parameter γ (see Methods) and quantified the joint effect of both loci on the disease

heritability by sibling relative risk (SRR) (see Appendix A.5). To ensure comparability across values of γ, we fixed the marginal relative risk at the locus of interest, and adjusted the marginal effect at the "remaining genome" locus to maintain SRR at 2, 4 and 8. We considered a locus of interest with f = 0.01 and set the marginal relative risk to 2.2 for models with no interaction ($\gamma = 1$) or synergistic interaction ($\gamma > 1$), and to 2.8 for models with antagonistic interaction ($\gamma < 1$) to illustrate the effect of antagonistic interaction with reasonable power. The qualitative impact of interaction on power was independent of these specific parameter choices (Results not shown).

Because the marginal effect at the locus of interest was constant, the power of the conventional case-control study was not affected by the considered interaction or by SRR. The power of TRAFIC increased with γ regardless of SRR across most interaction parameters considered (Figure 2-5). For synergistic interaction, the power rose quickly with γ; the exact trajectory depended on the SRR of the model. The power for models with a higher SRR increased faster for a lower γ, but the rate of increase also decreased faster for a higher SRR. Hence models with a lower SRR reached maximal power faster. In models of antagonistic interaction (γ < 1), TRAFIC rapidly lost power with decreasing γ. This loss of power was particularly pronounced for highly heritable diseases (SRR = 8). For SRR at 2, 4, and 8, TRAFIC was less powerful than the conventional design for γ < 0.52, 0.74, and 0.76, respectively (Figure 2-5a). However, the power started to increase when γ < 0.38, 0.31 and 0.26 for SRR=2, 4, and 8, respectively. For this extreme model of antagonistic interaction, a variant that was causal in a population sample had a protective effect in a family sample. Hence, the minor allele frequency on shared chromosome regions became lower than the minor allele frequency on non-shared chromosome

regions, generating power in a test for association.

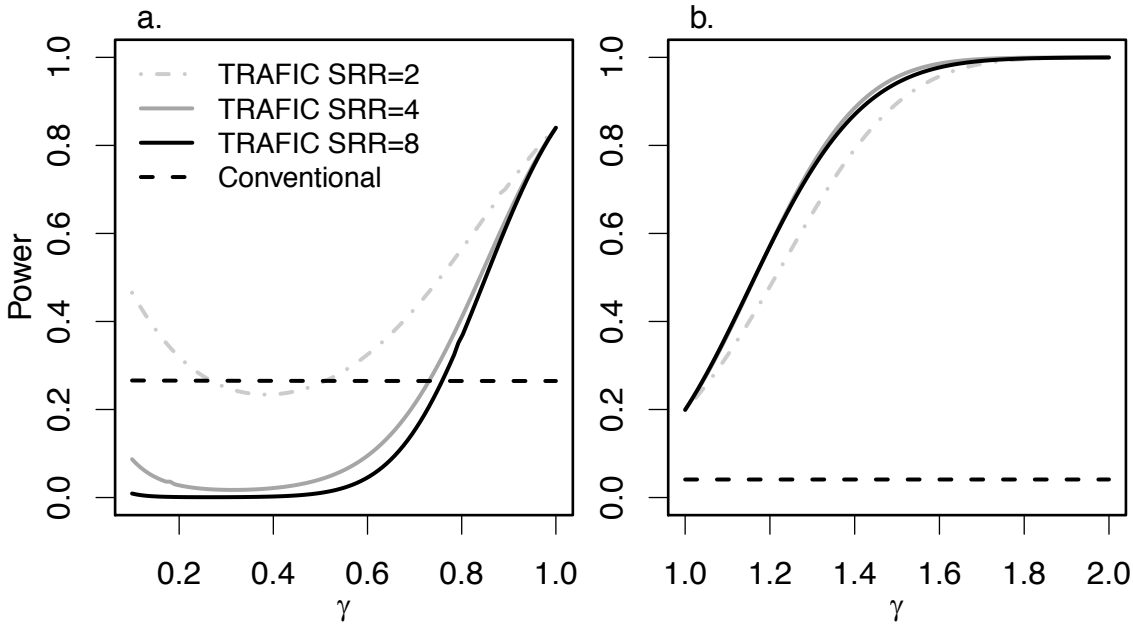


Figure 2-5. Analytical power of TRAFIC and the conventional case-control design under different models of gene-gene interaction. The horizontal axis displays the interaction parameter γ; gray and black lines represent different overall heritability parameterized as sibling relative risk (SRR). Panel a represents the result for antagonistic interaction ($\gamma < 1$); panel b represents the result of synergistic interaction ($\gamma > 1$). Results are shown for 2000 individuals (1000 sibpairs or 1000 cases and 1000 controls) at a significance level 2.5×10^{-6}.

2.4. Discussion

We introduce a new framework for gene-based association tests of rare variants leveraging affected sibpairs (TRAFIC). We compare the number of risk alleles located on chromosome regions shared IBD in an affected sibpair to the number of risk alleles located on chromosome regions that are not shared IBD. TRAFIC compares "cases" and "controls" within a sibpair as a matched design and is thus generally robust to population stratification. The test evaluates the null hypothesis of no association and can therefore generate a signal only in the presence of association and is powerful in the absence of linkage.

The proposed design of taking shared chromosome regions as new “cases” and non-

shared chromosome regions as new “controls” can be applied to any published gene-based tests. In this study, we evaluated the design for a collapsing gene-based test as the power of this test can be calculated without specifying minor allele frequency or effect size distribution of each risk variant, and it is therefore easier to obtain general conclusions. However, TRAFIC can also be applied to dispersion tests such as SKAT (Wu et al. 2011). We calculate the power of this new method using a general model for risk variants, which is specified by the summed allele frequency of risk variants, and mean and variance of relative risk for risk variants. We compared three study designs: (1) TRAFIC, (2) the conventional design of cases and controls, and (3) a design where cases are enriched for rare variants by selecting case individuals with affected relatives, assuming the same number of sequenced/genotyped samples. For diseases with prevalence ~1% and in the absence of gene-gene interaction, TRAFIC was more powerful than the conventional case-control design for variants with summed risk allele frequency less than 0.05, even though the conventional case-control design contained more independent observations. This power gain has two drivers. First, families ascertained to carry multiple affected individuals are more likely to segregate risk variants than random cases (Fingerlin, Boehnke, and Abecasis 2004; Zöllner 2012). Second, if such risk variants are rare, the founders of the pedigree are likely to only carry one copy. As the probability of carrying the risk variant is increased for each affected family member, this variant is more likely to be located on a shared chromosome. With increasing allelic heterogeneity, the probability for both affected siblings sharing an allele with a large effect size also rises, increasing the number of risk alleles located on shared IBD chromosome regions. Hence in the presence of allelic heterogeneity, the power of TRAFIC increased, while the power of the

conventional case-control design was unchanged.

The power of a family-based design also depends on the interaction between variants at the locus of interest and the remaining genome. Sampling from families with multiple affected individuals increases the overall genetic load for all cases. Hence, if the genetic effect at the locus of interest increases with overall genetic load, the power advantage of family-based designs over population-based designs is larger than under a model of no interaction. On the other hand, if the genetic effect of risk variants at the locus of interest decreases with overall genetic load, the power in family-based designs is smaller than the power under a model of no interaction and population-based designs can be more powerful. This effect has been described before for additive gene-gene interaction, which is a special case of genetic effect at the locus of interest decreasing with overall genetic load (Ionita-Laza and Ottman 2011; Zöllner 2012; Helbig, Hodge, and Ottman 2013).

Moreover, TRAFIC is generally robust to population stratification, as it compares IBD chromosome regions to non-shared chromosome regions in every sibpair thus naturally matching the genetic background of samples. This robustness can be violated in regions where one of the populations has a strong linkage signal while the other population has no evidence for linkage (Appendix A.6). However, this unlikely scenario only results in minor increase of the false positive rate and has thus little impact on the utility of our method. As the efficacy of current methods to control for population stratification in population based designs for rare variant tests is not clear (Mathieson and McVean 2012; Liu, Nicolae, and Chen 2013), family based designs may be necessary to avoid spurious association. TRAFIC achieves this robustness to stratification by using non-shared chromosome regions as controls at the cost of some reduction in power. As non-shared

chromosome regions have a higher risk allele frequency than chromosome regions in population controls, a test comparing shared chromosome regions against chromosome regions from unaffected controls may be more powerful than TRAFIC. However, such a design would be very susceptible to population stratification, even more than the selected cases design shown in Figure 2-4.

In conclusion, we have proposed TRAFIC using affected sibpairs for testing the association between a set of rare variants and the disease phenotype. TRAFIC is more powerful than the conventional case-control design under a wide range of models while being generally robust to population stratification.

Chapter 3
Powerful Statistical Testing for Rare Variant Associations with Binary Traits using Extended Families

3.1. Introduction

Genome-wide association studies (GWAS) have identified many variants associated with different complex diseases; however, the identified variants do not fully explain the estimated genetic contribution, leading to the search for missing heritability (Manolio et al. 2009; Eichler et al. 2010). Rare variants are hypothesized to explain some of the missing genetic contribution (Manolio et al. 2009; Eichler et al. 2010). Importantly, these rare variants are abundant in human populations. In a sequencing study of 202 drug-targeted genes from 14,002 individuals, more than 95% of single-nucleotide variants found had minor allele frequency (MAF) < 0.5% (Nelson et al. 2012). In addition, rare variants can make important contributions to complex disease risk. In a recent sequencing study of 63 known prostate cancer risk regions, variants with minor allele frequency 0.1-1% accounted for 42% of genetic risk (Mancuso et al. 2016).

Using conventional approaches to investigate the role of rare variants in complex diseases has limited statistical power. For example, to study the association between rare

variants and complex diseases, conventional case-control studies require huge sample sizes (>10,000 individuals) to have adequate power (Manolio et al. 2009; Nelson et al. 2012). In contrast, family samples can provide power gain for association studies, especially for rare variants (Feng et al. 2015). When combined with ascertainment strategies, family-based methods further benefit from the enrichment of rare variants. For example, collecting cases with ascertained siblings and comparing them to population controls can significantly increase the power relative to a population case-control design (Zöllner 2012).

Another advantage of family studies arises in the context of population structure, which is a common confounder in association studies (Price et al. 2006; Zawistowski et al. 2014). Rare variants are often population-specific or shared by closely related populations (The 1000 Genomes Project Consortium 2015); thus rare variants are more prone to population stratification than common variants. Even minor population stratification can significantly inflate type I error and result in spurious findings (Zawistowski et al. 2014). Although many methods have been developed to correct for population stratification for common variants in population samples, it is not evident that these methods are as effective when applied to rare variants (Y. Zhang, Shen, and Pan 2013). Family data, in which family members are typically from the same genetic and environmental background, can protect against spurious signals caused by population stratification and other environmental confounders (Ott, Kamatani, and Lathrop 2011). These advantages of family studies have motivated adaptations of many population-based methods for rare variant associations applied to family data (H. Chen, Meigs, and Dupuis 2013; X. Wang et al. 2013; Schaid et al. 2013; Q. Zhang et al. 2014).

To increase power, combining the signal across multiple rare variants in a gene has been shown to be effective in identifying associations between rare variants and complex diseases (B. Li and Leal 2008; Zawistowski et al. 2010; Wu et al. 2011). These gene-based tests have facilitated association studies and identified many associated rare variants (Lee et al. 2014). Extending such methods to family samples requires modeling the relatedness in the samples; otherwise, the standard error of the test statistic is underestimated, leading to inflated type I error. To correct the underestimated standard error, a popular method is to incorporate the kinship matrix in the mixed model to account for relatedness (Kang et al. 2010; Schaid et al. 2013). These corrections decrease the effective sample size, so that family-based approaches often appear to have inferior power compared to similarly-sized population case-control studies. An alternative approach to account for relatedness is to use methods that only rely on within-family information such as FBAT (Rakovski et al. 2007); De et al. (2013) and Ionia et.al. (2013) extended FBAT to gene-based tests. This class of within-family approach often requires separating a large family into many parent-offspring trios, leading to power loss (Laird and Lange 2006). In general, although the approaches that consider only within-family information are robust to population stratification, they are not as powerful as the approaches adapted from the population-based design, which exploit both between- and within-family information.

Recent methods for rare variant association in families exploit that rare risk variants are more likely to reside on chromosomes shared identity-by-descent (IBD) among affected family members. Hence, comparing the distribution of rare variants in IBD chromosomes and non-IBD chromosomes in affected sibpairs can be more powerful than a conventional case-control study (Epstein et al. 2015; Lin and Zöllner 2015). Here we

propose testing for rare variant associations in pedigree studies (TRAP) that compares the distribution of rare variants in chromosomes shared among affected family members to other chromosomes. A variant that increases disease risk is more likely to be transmitted from founders to affected offspring thus creating a departure from the Mendelian segregation (Figure 3-1). TRAP is applicable to any type of pedigree structures, from sibpairs to multi-generation families. We evaluate this approach using simulations with realistic parameter configurations and show that, given the same number of sequenced individuals, TRAP can substantially outperform the conventional population-based test and existing family-based methods.

3.2. Methods

3.2.1. Overview

First, we introduce our proposed family-based test, TRAP. Second, since TRAP requires founder information, which may not be accessible for every family, we devise an imputation algorithm to account for missing founders. Third, we describe a disease model that accounts for the heritability in families and elaborate how we conduct simulations based on this disease model.

3.2.2. Test for rare variant associations in pedigree studies (TRAP)

Consider a set of fully genotyped families of arbitrary size and structure with multiple affected individuals. Moreover, assume the chromosome inheritance vector is known for each family. In practice, several methods can provide a precise estimation of inheritance vectors with high certainty (Abecasis et al. 2002). In the following, we aim to jointly test

all putative risk variants in a region of interest (e.g. a gene) for associations with the disease phenotype. To account for co-segregating rare variants in a region, we define a carrier chromosome as a region carrying at least one putative risk variant. If any rare variant is associated with the disease, we expect affected family members to share carrier chromosomes more than expected at random. Analogously, we expect a depletion of carrier chromosomes among unaffected family members. To test for this pattern, we define $X_{ij} \in \{0,1,2\}$ as the number of carrier chromosomes for the j^{th} member in i^{th} family. We evaluate the distribution of carrier chromosomes in each family by calculating $T_i = \sum_j X_{ij} Y_{ij}$, where $Y_{ij} = 1$ if the j^{th} member in i^{th} family is affected and $Y_i = -1$ otherwise. Next, we compare the observed T_i to the null expectation that carrier status is unrelated to the disease risk.

To carry out this comparison, we condition on the inheritance vector in each family (Figure 3-1) and on the number of carrier chromosomes among founders. For example, a nuclear family has four founder chromosomes, each with its own inheritance path. If there is one carrier chromosome in the founders, this carrier chromosome can thus take one of four possible inheritance paths. Under the null hypothesis, each founder chromosome is equally likely to be a carrier chromosome. Under alternative that a variant increases disease risk, carrier chromosomes are more likely to be shared among affected family members and to be transmitted through paths that include multiple affected individuals. Conditional on the inheritance vector in family *i*, we can thus generate the null distribution of T_i and calculate $\mu_i = \frac{1}{n_c}\sum_{k=1}^{n_c} T_i^{(k)}$ and $\sigma_i^2 = \frac{1}{n_c}\sum_{k=1}^{n_c}\left[T_i^{(k)} - \mu_i\right]^2$ where n_c is the number of paths and $T_i^{(k)}$ is the count of affected individuals inheriting a carrier chromosome through a particular path (see Figure 3-1 for example where $\mu_i = 1$ and

$\sigma_i^2 = 1.5$). We then calculate a joint test statistics T across families,

$$T = \frac{1}{S_n}\sum_{i=1}^{n}(T_i - \mu_i)$$

where $S_n^2 = \sum_{i=1}^{n}\sigma_i^2$. Using the Lyapunov Central Limit Theorem, we can show that the test statistic T follows a standard normal distribution (see Appendix B.1 for details). Based on the resulting z-score, we test for the association. Note that $\sigma_i = 0$ means that, in family *i*, either no founder carries a rare risk variant or that all founders carry rare risk variants on both chromosomes. Such families do not contribute to T, and we do not include them in the final test statistic.

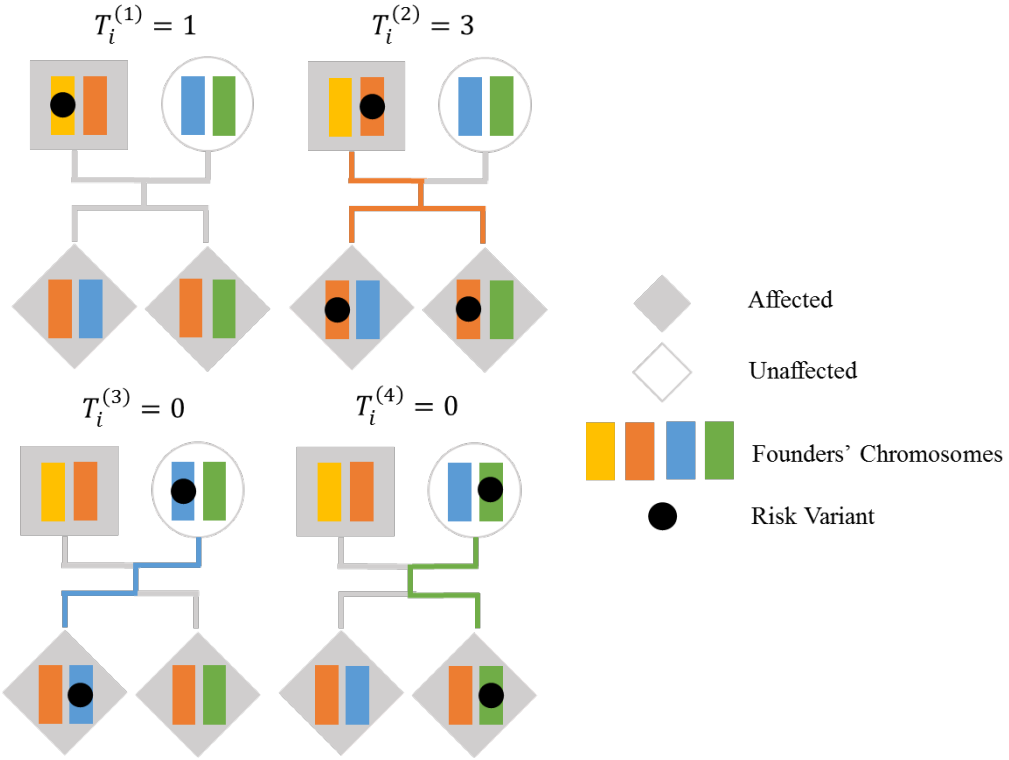


Figure 3-1. Illustration of counting carrier chromosomes for a family with two founders and two offspring assuming a single variant in the region. There are four possible paths to place the variant in founder chromosomes. The colored-line indicates the path that a variant is transmitted to offspring. $T_i^{(k)}$ is the count of carrier chromosomes given that carrier chromosome is transmitted through a particular path. Conditional on the inheritance vector in family i, we can thus enumerate all four possible $T_i^{(k)}$. Under the null hypothesis that each placement is equally likely to occur, we then can calculate $\mu_i = \frac{1}{4}(1+3+0+0) = 1$ and $\sigma_i^2 = \frac{1}{4}[(1-1)^2 + (3-1)^2 + (0-1)^2 + (0-1)^2] = 1.5$. Under alternative that a variant increases the risk of developing disease, the variant is more likely to appear on the second (orange) founder chromosome and reach two affected offspring.

3.2.3. Imputation algorithm for missing founders

TRAP depends on information from founders to perform the association test. In practice, this information may be missing, especially for larger pedigrees or late onset diseases. Here, we devise an algorithm to impute founder genotypes and inheritance vector. There are two steps in the algorithm. The first step is to impute inheritance vectors and the phasing of variants. Given the number of IBD chromosomes shared between family members, we determine an inheritance vector for each family. Then, given the inheritance vector, we determine the phasing configuration based on the observed genotype for each family member. If there is an ambiguous phasing, we randomly draw a possible phasing based on its relative likelihood (see Appendix B.2.1 for details). From the above procedure, we can identify the number of transmitted founder chromosomes and their carrier status. The second step is to impute the carrier status for the non-transmitted founder chromosomes. We use the probability of a founder chromosome being a carrier $p^{carrier}$ and thus impute the number of carrier chromosomes in missing founders. $p^{carrier}$ is estimated based on all observed transmitted founder chromosomes across families. Alternatively, $p^{carrier}$ can be obtained from existing population variant databases, e.g. 1000 Genomes Project which contains the majority (> 95%) of variants with allele frequency > 0.5% (The 1000 Genomes Project Consortium 2015). Specifically, for a family with $n_i^{missing}$ missing founder chromosomes, we impute the number of carrier chromosomes $x_i^{carrier}$ following a binomial distribution with probability $p^{carrier}$, $x_i^{carrier} \sim Binomial\left(n_i^{missing}, p^{carrier}\right)$. For example, for a sibpair family where siblings share both chromosomes, there are $n_i^{missing} = 2$ missing founder chromosomes and we impute the carrier status for missing founder chromosomes $x_i^{carrier} \in \{0,1,2\}$. Next, we

use imputed samples to perform TRAP and apply multiple imputation when generating test statistics to account for the imputation uncertainty (details in Appendix B.2.2).

3.2.4. Simulation design

Using simulations, we compare power between TRAP and two existing family-based methods, Pedgene which uses both between- and within-family information by comparing cases to controls and adjusting for relatedness in families (Schaid et al. 2013), and FB-SKAT which extends the within-family-approach FBAT to gene-based test (Ionita-Laza et al. 2013).

We first describe a disease model for population samples and later extend to family samples. Here we define p as the disease risk, $\underline{G} = \{g_1, g_2, \ldots, g_m \,|g_i \in (0,1,2)\}$ as an indicator vector for carrying the variant allele at m^{th} variant of interest in the region. For cases and controls from a population, we model the disease risk as $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \boldsymbol{G} + \epsilon$, where β_0 is the prevalence, β_1 is the effect size of the risk variant assuming all risk variants have the same effect size, and $\epsilon \sim N(0, \sigma_E^2 = 0.5)$ represents the random error. We generate an equal number of cases and controls, and apply a simple burden test collapsing variants in a region into a carrier chromosome (B. Li and Leal 2008).

To generate families, we model the heritability in a family by incorporating the kinship matrix of i^{th} family $\boldsymbol{K}_i$ and modify the disease model as $\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 \underline{G}_{ij} + F_{ij} + \varepsilon_{ij}$. Here, *ij* denote the j^{th} individual in the i^{th} family, and $F_{ij} \sim N(0, \sigma_{PE}^2 = 0.5)$ is the shared polygenic and environmental effect in a family. The covariance matrix in i^{th} family is $\boldsymbol{V}_i = 2\sigma_{PE}^2 \boldsymbol{K_i} + \sigma_E^2 \boldsymbol{I}$. Under the null hypothesis, we adjust β_0 so that the diseases prevalence is 10%, unless specified otherwise.

We simulate sequence data using a coalescent-model-based simulator COSI (Schaffner et al. 2005) and generate ten thousand 1kb chromosomes. We randomly pick variants whose frequency sums to *f* and fix these variants as causal in the simulations. To generate $\underline{G_{ij}}$, assuming no recombination in families, we sample two chromosomes for each founder from those ten thousand chromosomes; then, each offspring inherits two chromosomes following Mendel law (MacCluer et al. 1986). Given the generated chromosomes, we simulate the phenotype for all family members based on the disease model introduced above and only keep the families that match the specified number of affected and unaffected in a given family structure. To generate families with a given pedigree structure, we simulate family by family until there are 1000 families generated. Then we apply TRAP, Pedgene, and FB-SKAT to compare power. In the following sections, we use 1,000,000 replications to evaluate type I error rate and 1,000 replications for each scenario to calculate power.

3.3. Results

Using simulations with regions of 1kb, we compared the proposed TRAP with Pedgene (Schaid et al. 2013), FB-SKAT (Ionita-Laza et al. 2013), and a conventional case-control design. Results that applied TRAP to all affected and unaffected family members had a negligible difference in power compared to the affected-member-only analysis (Figure B-5). Thus, in the following, we consider TRAP applied to founders and affected family members only, leading to lower sequencing cost. First, we demonstrate all methods are calibrated under the null hypothesis. Second, we compare power across family-based methods and contrast them to a conventional case-control design. Third, we explore the

power change in different pedigree structures, specifically to see how the power changes with the increasing pedigree size and ascertainment. Lastly, we present the results using imputation to account for missing founders.

3.3.1. Type I error rate

To evaluate type I error rate, we simulated 1000 two-generation families and three-generation families under the null. We assumed inheritance vectors were known in this evaluation. We considered the variants with population carrier chromosome frequency of 0.01. To limit computational burden, we set type I error rate at 2.5×10^{-4}. Table 3-1 shows all methods retained nominal type I error rate across considered family structures.

	TRAP	Pedgene	FB-SKAT
Two-generation family-- 2 affected 2 unaffected	2.0×10^{-4}	2.5×10^{-4}	2.2×10^{-4}
Two-generation family-- 3 affected 1 unaffected	2.1×10^{-4}	2.0×10^{-4}	1.9×10^{-4}
Three-generation family-- 2 affected 5 unaffected	2.5×10^{-4}	2.4×10^{-4}	2.6×10^{-4}

Table 3-1. Type I error rate evaluation for carrier chromosome frequency 0.01 and different family structures with nominal $\alpha = 2.5\times10^{-4}$.

3.3.2. Power comparison in nuclear families

We compared TRAP to existing methods, Pedgene and FB-SKAT. Pedgene extended the existing methods for population samples to family data, and FB-SKAT extended the framework of FBAT to gene-based test. Both methods can choose to perform gene-based test in “burden” or “variance component” style. Since the simulations in this study considered only risk variants, which resulted in favoring burden tests, we only presented

those results in burden style in the following sections. We considered 1000 nuclear families with two affected and two unaffected, and three disease models under population carrier chromosome frequency f of 0.01, 0.05 and 0.2. We calculated the power over a range of odds ratios, r. Among the considered family-based tests, assuming disease prevalence of 1%, TRAP was uniformly the most powerful method from rare to common variants followed by Pedgene. FB-SKAT, which only relied on within-family information, was as expected to be the least powerful method. For rare variants with f = 0.01 and r = 2.5, the power for TRAP, Pedgene, and FB-SKAT was 0.767, 0.423, and 0.043, respectively. The advantage in power for TRAP maintained for variants with f = 0.20; at r = 1.4, the power was 0.908, 0.740, and 0.004, respectively.

When comparing to a conventional case-control design assuming a simply burden test (B. Li and Leal 2008), using an equally-sized population samples was more powerful than all considered family-based tests, except when f = 0.01, TRAP was the only method being more powerful than the population case-control design. In contrast, for higher disease prevalence at 10%, none of the family-based methods showed an advantage in any scenario, except only TRAP had a comparable power to the conventional case-control design for f = 0.01 (Figure B-6).

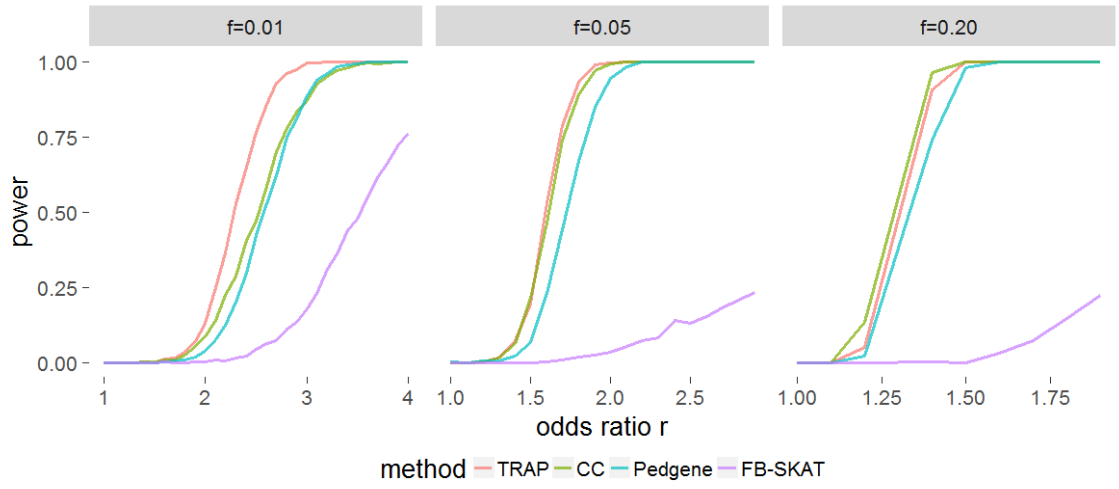


Figure 3-2. The power comparison evaluated at $\alpha = 2.5\times10^{-6}$ and prevalence 1% across TRAP, Pedgene and FB-SKAT. Power curve is shown as a function of effect size r (odds ratio) of risk variants in a gene with carrier chromosome frequency f = 0.01, 0.05, 0.20.

3.3.3. Power comparison with additional generation or affected members in families

To evaluate the effect of pedigree size and ascertainment on TRAP, we considered multiple pedigree structures with different levels of ascertainment. We evaluated four pedigree structures and coded each structure by the number of generations (g), number of affected individuals (a) and the number of unaffected individuals (u); for example, a two-generation family with three affected and one unaffected was denoted as 2g.3a.1u; we allowed any family member to have the disease (Figure B-7 shows the illustration of different pedigree structures). Assuming we can reliably impute offspring genotypes given sequenced founders in a family and the cost of genotyping array was relatively negligible compared to the cost of sequencing, we thus considered the same number of sequenced founders for different family structures, resulting in 1000, 667, and 286 families for 2g.3a.1u, 3g.3a.4u, and 3g.3a.19u (3g.5a.17u). In general, these pedigree structures differ

in the number of family members (4, 7, and, 22 members), the number of affected (3 and 5 affected) and the number of generations (2 generations vs. 3 generations).

Figure 3-3 shows the power comparison across pedigree structures for TRAP. For f = 0.01, given three affected family members across pedigree structures, TRAP with two-generation pedigree 2g.3a.1u had a similar power compared to three-generation pedigree 3g.3a.4u and 3g.5a.17u. For instance, for r = 2.5, TRAP with 2g.3a.1u, 3g.3a.4u, and 3g.5a.17u had power of 0.546, 0.529 and 0.460, respectively. Note that pedigrees with a higher proportion of affected individuals were more ascertained with the disease; therefore, these families were more likely to carry risk variants, resulting in power gain. The level of ascertainment can be quantified by how challenging to collect such an ascertained family. To collect a family with 2g.3a.1u, 3g.3a.4u, 3g.5a.17u and 3g.3a.19u ascertainment, the chance of gathering one such family were one in 201, 37, 20 and 5 equally-sized families, respectively. Thus, 3g.3a.17u was the least ascertained pedigree and thereby TRAP had the least power with. Although 2g.3a.1u was the most ascertained pedigree (also the most challenging to collect), it only achieved a comparable power as much-less-ascertained 3g.3a.4u, and 3g.5a.17u, suggesting a pedigree with three generations was more informative to TRAP than two generations. In comparison to existing family-based methods, TRAP was more powerful than Pedgene and FB-SKAT in all considered pedigree structures (Figure B-8).

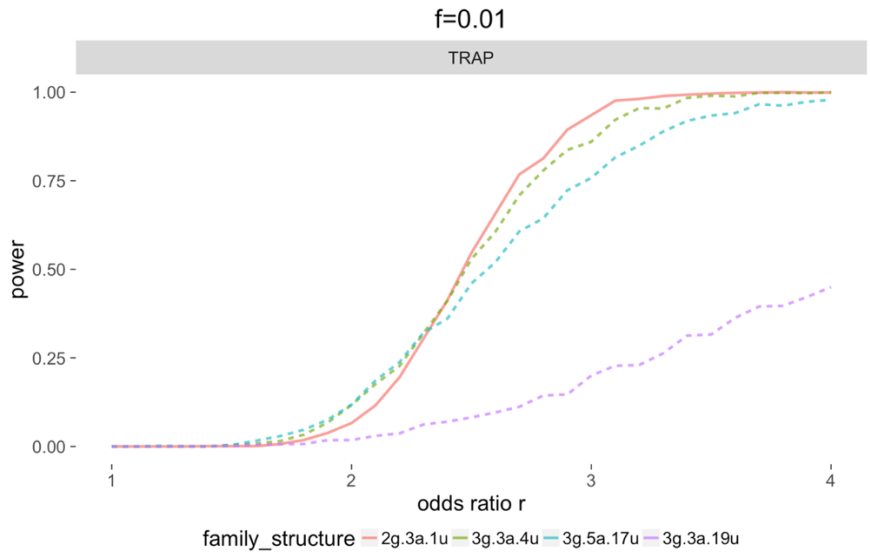


Figure 3-3. Power curve across pedigree structures evaluated at $\alpha = 2.5\times10^{-6}$ and f = 0.01 as a function of odds ratio of risk variants for TRAP. Solid lines represent two-generation families and dashed lines represent three-generation families

3.3.4. Power study with missing founders

Performing TRAP depended on founder information. To apply TRAP to families with missing founders, we developed an imputation procedure. Specifically, we imputed the inheritance vector and founder carrier chromosomes (more details in Appendix B.2). To evaluate TRAP performance using imputed samples, here we generated 1000 families with two founders and three siblings, in which two siblings were affected and one was unaffected in order to compare with Pedgene, which is not applicable to affected-only design. We considered four scenarios in which all founders were missing or a random subset of founders were missing in 20%, 50%, and 80% of families. We compared TRAP imputation to TRAP with compete founder information (but did not consider founder phenotypes) and to Pedgene, which did not require information from founders. Since FB-SKAT could not perform without founders and did not implement an imputation procedure in their program, it was excluded in this study. Type I error rate was set at 0.05

to demonstrate the change and there was no inflation given a more stringent level (Table B-1).

	f=0.01	f=0.20
Complete data -- No imputation	0.056	0.041
Imputed samples with 20% missing founders	0.047	0.036
Imputed samples with 50% missing founders	0.021	0.014
Imputed samples with 80% missing founders	0.011	0.007
Imputed samples with 100% missing founders	0.008	0.005

Table 3-2. Type I error rate evaluation for TRAP with and without imputed samples under different population carrier frequency given $\alpha = 0.05$.

First, assuming all founders were missing, we evaluated if TRAP with imputed families was a valid test. For rare variants with *f* = 0.01, TRAP became conservative with type I error rate deflated from 0.056 without imputation to 0.008 using imputation (Table 3-2). Next, we evaluated the degree of power loss due to being conservative and using imputed families in Figure 3-4; since all founders were missing, this scenario served as a lower bound of how TRAP imputation resulted in power loss. For *r* = 2, TRAP imputation had reduced power from 0.799 assuming known founders to 0.413. However, TRAP imputation had comparable power to Pedgene; given sufficient large effect size, e.g. *r* = 2.3, TRAP imputation have an advantage in power over Pedgene, 0.703 to 0.655. Note that there were two cases for every one control in Pedgene, and this was not optimal in a case-control design; thus, increasing the number of controls in families would increase the power for Pedgene.

As for common variants *f* = 0.20, TRAP imputation also had deflated type I error rate (0.005) and resulted in a similar power loss as compared to rare variants (*f* = 0.01). Using the carrier chromosome frequency from external databases was the most efficient way to

impute missing founders and resulted in only minor loss in power. This approach, however, was highly sensitive to the bias to the true carrier chromosome frequency; even a small bias could lead to substantially inflated type I error (Figure B-9).

Lastly, we considered the degree of power loss when there were some observed founders in the samples. With these additional founders, we can improve the estimate for carrier chromosome frequency, leading to less power loss. Figure 3-4 shows that when there were only 80% of families with missing founders, the power was improved relative to all families with missing founders. For example, for f = 0.01 and r = 2, the power for TRAP imputation is 0.518 compared to 0.413 when all founders were missing. When there were only 20% of families with missing founders, the power loss was negligible compared to no-missing-founder analysis (with power 0.761 and 0.799).

In sum, from rare to common variants with ascertained families, TRAP was more powerful than the two existing family-based methods under many pedigree structures considered. Besides, for disease prevalence 1% and rare variants, TRAP outperformed the population case-control design. Across pedigree structures, families with more generations and more affected family members increased power. Importantly, we developed the imputation algorithm for families with missing founders. Even when all founders were missing, TRAP using imputation was comparable to existing methods and the loss in power decreased with the increasing number of observed founders.

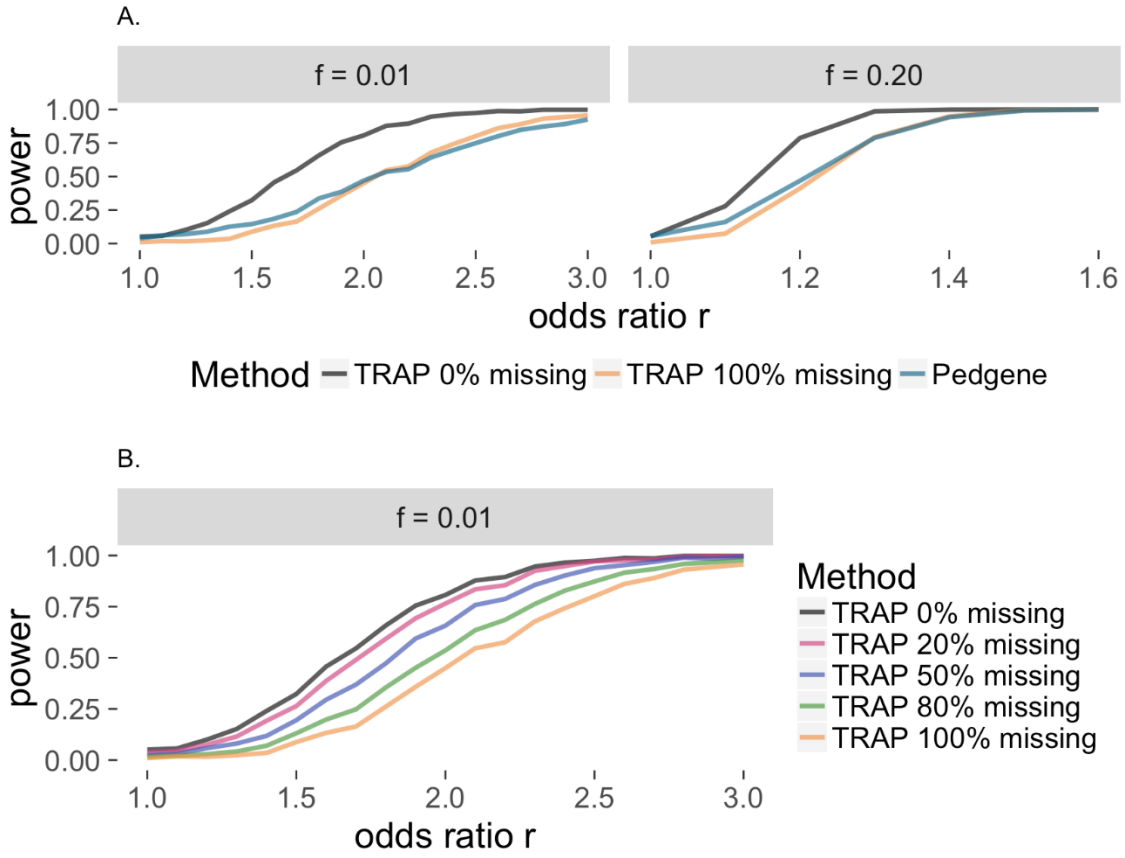


Figure 3-4. Power curve for TRAP using imputaion for differenct proportion of missing founders evlaued at $\alpha = 0.05$. The imputation algorithm estimates carrier chromosome frequency based on observed samples. Panel (A) shows the results for TRAP with complete data and 100% missing founder compared to Pedgene. Panel (B) shows the result for TRAP under different proportion of families with missing founders.

3.4. Discussion

We propose a family-based method TRAP designed for rare variant association studies using extended families. TRAP exploits that risk variants are more likely to be shared IBD among affected family members. To leverage this observation, TRAP compares the number of carrier chromosomes in a family to the expectation conditional on the founder genotypes and inheritance vectors. Equivalently, we propose to test if the variant is equally likely to be transmitted through all possible transmission paths from founders to offspring. Based on simulations with realistic parameter settings, TRAP was uniformly

more powerful than existing family-based methods in all considered pedigree structures; Given disease prevalence 1% and rare variants with $f = 0.01$, TRAP also outperformed the conventional case-control study.

Better using IBD information can increase the power to detect rare variant associations as shown in affected-sibling designs (Epstein et al. 2015; Lin and Zöllner 2015). The proposed method TRAP not only better exploit IBD information but also provide the flexibility to apply to arbitrary pedigree structures, from sibpairs to multi-generation families. Studying large and extensive pedigree, i.e. more founders in a family, increases the chance of observing a segregating risk variant in the family, leading to power gain. In addition, with the devised imputation algorithm, TRAP can also apply to families where the founder genotype is not available with minimal power loss. This imputation algorithm can easily extend to scenarios in which some of family members have missing genotypes but with phenotypic information available. Specifically, the missing genotype can be replaced by the expectation conditional on available relatives (W.-M. Chen and Abecasis 2007). Since acquiring disease status is more accessible than gathering genotype information, we expect TRAP is especially useful when applying to extensive pedigrees with imputation.

As TRAP exploits within-family information and includes founders in the test, it achieves greater power compared to the existing family-based method Pedgene, which uses both between- and within family information and the conventional case-control design. The reason is that, for a family with multiple affected family members and low disease prevalence, it is more likely that founders carry the risk variant and this influences TRAP in two ways. First, the number of informative families increases, i.e. at least one

founder carries the risk variant. Second, the rare risk variant is more likely to be shared among affected and TRAP efficiently uses the inheritance vector to test for this association. These two conditions significantly increase the power of TRAP to detect rare variant associations compared to the conventional case-control designs. Although the allele count of rare risk variants also increases for Pedgene, it does not directly use the sharing of rare risk variant between affected family members, leading to less power gain than TRAP. Compared to FB-SKAT that only uses within-family information by separating an extended family into many trios, TRAP gains additional power by taking into account founders' affected status and jointly considering the information across many affected family members.

In this study, we evaluated statistical power between TRAP and the conventional case-control design assuming the same number of sequenced individuals without considering the increased cost of recruiting ascertained families. Note that the power gain for TRAP relative to the case-control design depended on the level of ascertainment in sampled families. For nuclear families with two affected and two unaffected family members, TRAP was only advantageous to identifying rare variant associations for low prevalence diseases. However, TRAP allows the design which collects only founders and affected individual in families, and has a negligible power loss; hence, given fixed sequencing costs, using TRAP allows to sample more families and thus achieve higher power. In contrast, most existing family-based methods, such as Pedgene, are not applicable to affected-only studies.

Given the decreasing cost of sequencing technology, researches are bound to gather a large number of samples; thus, population stratification is inevitable to be a challenge

for large-scale association studies. Since TRAP evaluates the disproportionate sharing among affected family members within each family, TRAP is robust to population stratification when families come from different ancestral backgrounds. However, when the founders within a family are heterogeneous in origin, TRAP is subject to population stratification by assuming these founders have the same carrier chromosome frequency. As a future improvement, we can infer founders' ancestry and their individual-specific carrier chromosome frequency (Conomos et al. 2016), and then accommodate TRAP to mitigate the confounding effect of population stratification.

TRAP requires inheritance vectors as input and we assume inheritance vectors are known in this study. For nuclear families, multiple software can reliably infer inheritance vectors (X. Li and Li 2011; Roach et al. 2011; O'Connell et al. 2014). However, the inference of inheritance vectors for large pedigrees can be challenging, especially when there are multiple ungenotyped members/founders in the pedigree. In addition, large pedigrees are more prone to genotyping error which influences the accuracy of inferred inheritance vectors than nuclear families (X. Li and Li 2011). As the size of pedigree grows, the computational time for inferring inheritance vectors increases with multiple possible inheritance vectors for a given pedigree. For such a large pedigree, we can extend TRAP to evaluate associations based on the dosage of each inferred inheritance vector; though, this also requires additional computational time. Future work is required to evaluate the performance of TRAP on large pedigrees accounting for the uncertainty of inferring their inheritance vectors.

Overall, family designs are especially informative for rare variant association studies. We develop a family-based method that can both apply to arbitrary pedigree structures

and efficiently exploit the information of rare variants shared on IBD chromosomes. We expect TRAP is particularly useful for family studies with ascertainment, especially for multi-generation pedigrees.

Chapter 4
Increasing Power for Testing Rare Variant Associations with Continuous Traits using Extended Families

4.1. Introduction

Genome-wide association studies have identified many common variants that account for a modest portion of heritability for a given disease (Eichler et al. 2010). To further explore the genetic contribution to the complex diseases, rare variants are hypothesized to explain some of the remaining heritability (Manolio et al. 2009). In fact, rare variants are abundant in human populations (The 1000 Genomes Project Consortium 2015). Thus, many studies have focused on identifying rare variants and their contributions to complex diseases (Cirulli and Goldstein 2010).

The challenge of using population samples to study rare variants is the low statistical power to identify the rare variants associated with a trait (Lee et al. 2014). To improve power to identify associated rare variants, one popular strategy is to aggregate the rare variants in a gene region to jointly test for associations in "burden" or "variance-component" styles (B. Li and Leal 2008; Wu et al. 2011). An alternative strategy is to use

family samples. Extending gene-based tests to family-based samples can further improve power since genetic loading at susceptible loci can be enriched in families (Ott, Kamatani, and Lathrop 2011). These extensions can be categorized into two classes: to adjust for the relatedness in families, the first class of methods extend population gene-based tests by employing generalized estimating equation or mixed effect models (H. Chen, Meigs, and Dupuis 2013; Schifano et al. 2012; Schaid et al. 2013; X. Wang et al. 2013). Although valid, this adjustment typically reduces the effective sample size, offsetting the benefit of the increased loading at susceptible loci in the family; this makes these family gene-based tests fall behind their counterparts for population samples. The second class of methods is based on FBAT (Laird, Horvath, and Xu 2000), that uses within-family information to test for the equilibrium-transmission of alleles in the family, and extends to gene-based test; these methods are robust to the confounding effect due to population stratification (De et al. 2013; Ionita-Laza et al. 2013). Instead of only using within-family information, Jiang, Conneely, and Epstein [2014] proposed to use between-family information to select the variants to include in the within-family association test to reduce the multiple-testing burden. Fang, Sha, and Zhang [2012] included the between-family information and formed a combined test; a combined test is more powerful than using only within-family information, however, losing its immunity to population stratification. However, FBAT-based method often breaks an extended family into many trios where only heterozygous parents contribute to the test. For rare variants, most parents would be homozygous. This significantly decreases the number of informative transmissions; thus, these methods do not efficiently use the within-family information.

In Chapter 3, TRAP presents a new framework to exploit within-family information by

including all homozygous parents in a family in which at least one parent is heterozygous. By directly considering the sharing of rare variants among affected family members, TRAP improves power for detecting rare variants associated binary traits as compared to FB-SKAT, which is a gene-based test extended from FBAT (Ionita-Laza et al. 2013) and Pedgene, which is an extension of population gene-based test (Schaid et al. 2013). By using a similar idea, we can extend this efficient within-family test framework to studying continuous traits. Continuous traits are routinely collected in many disease studies. For example, in a diabetes study, patients are usually measured for their lipid, glucose level, waist/height ratio, and blood pressure. Therefore, the goal of this chapter is to propose a new efficient gene-based method to use family data to increase power for the identification of associated rare variants, focusing on continuous traits.

To increase power based on the fact that, given a rare variant increases the trait, two family members who have a high-level trait are more likely to share this rare variant identity-by-descent (IBD). Collecting ascertained families, which have many family members who have a high-level trait, can increase the observed frequency of risk alleles, leading to power gain. Although powerful, the ascertainment requirement also increases the cost of collecting family samples, resulting in a smaller sample size (given a fixed budget), which can offset the power gain. On the other hand, by collecting a larger sample of unascertained families, i.e. randomly selected families in a population, may still allow us to observe excessive copies of risk alleles within a family, described as the 'Jackpot' effect (Feng et al. 2015). Thus, developing a method that has adequate statistical performance for both scenarios can ease the process of finding rare-variant associations and not have to consider whether the family is "ascertained" or not.

Here, we propose test for rare-variant associations with continuous traits using extended families (TRACE), which can be applied to both non-ascertained and ascertained scenarios. Intuitively, within each family, we test if the family members with a similar trait are more likely to share the IBD variant. Equivalently, if a variant increases the trait value, this variant would pass through the founder chromosome that can reach many family members with high trait values (as illustrated in Figure 4-1). Therefore, using within-family information, we test if the transmission path is associated with the trait of family members. Moreover, within- and between-family information provide independent information about association signals (Fang, Sha, and Zhang 2012). To further increase power, we design TRACE as a combined method that can use both within- and between-family information.

In the following, we first describe and explain how to use within-family to perform the test. Second, we show how TRACE combines within- and between-family information to increase power. Using simulations, we show that under a non-ascertained scenario, TRACE is more powerful than the existing family-based methods for large pedigree, but less powerful for small and medium pedigree; for ascertained scenarios, TRACE has an advantage over the existing approaches in all considered scenarios.

4.2. Method

In this section, we first introduce TRACE_W using within-family information. Then, we introduce TRACE using both within- and between-family information. Finally, we explain the simulation model that accounts for the relatedness and heritability in families to evaluate the proposed and existing methods under non-ascertained and ascertained

scenarios.

4.2.1. Test for associations based on within-family information: TRACE_W

Considering a region in the human genome, we jointly test if rare variants are shared by the family members who have a similar trait value. In a chromosome region, we use a simple collapsing approach (B. Li and Leal 2008) to collapse the variants of interest, e.g. nonsynonymous variants, and define the carrier chromosome as a chromosome region carrying at least one variant of interest. For j^{th} individual in i^{th} family, let $x_{ij} \in \{0,1,2\}$ be the number of carrier chromosomes for every individual and Y_{ij} be the trait value, $i = 1 \dots n_{\mathrm{fam}}$ and $j = 1 \dots n_{\mathrm{i}}$. In addition, we center the individual's trait by subtracting the family mean $\overline{Y_i}$. To quantify the covariance between carrier chromosomes and the trait, we calculate $T_i = \sum_j^{n_i} (Y_{ij} - \overline{Y_i})\, X_{ij}$. We can then use T_i to test for the association. Intuitively, if the variant is associated with the trait, T_i would deviate from its expectation $E(T_i)$ assuming no association. We then calculate the expectation for every family and aggregate the information across families to test for the association between the variants and the continuous trait.

To calculate $E(T_i)$, we use the inheritance vector; the inheritance vector indicates possible paths for a variant to be transmitted from founders to offspring in a family (as shown in Figure 4-1). For example, for a nuclear family, two parents provide four possible transmission paths. Under the null hypothesis that the variants are not associated with the trait, the carrier chromosome is equally likely to be transmitted through any transmission path from founders to offspring. Given the founder genotypes L_i and the inheritance vector IV_i, we can thus enumerate all possible transmission paths *c* and

calculate the expectation $E(T_i) = \sum_{c|J_i,IV_i}^{n_c} \sum_j^{n_i} (Y_{ij} - \overline{Y_\iota}) X_{ij|c} / n_c$, where n_c is the number of possible transmission paths. The test statistic using within-family information, $T_W = \sum_i^{n_{\text{fam}}} [T_i - E(T_i)]$, aggregates the information across families. To derive the null distribution of T_W, we can rewrite within-family information, $T_i - E(T_i)$, as a summation of the excessive transmission of carrier chromosome and the trait of each family member as below,

$$T_i - E(T_i) = \sum_j^{n_i} (Y_{ij} - \overline{Y_\iota}) X_{ij} - \sum_{c|L_i,IV_i}^{n_c} \sum_j^{n_i} (Y_{ij} - \overline{Y_\iota}) (X_{ij|c} / n_c)$$

$$= \sum_j^{n_i} (Y_{ij} - \overline{Y_\iota}) X_{ij} - \sum_j^{n_i} (Y_{ij} - \overline{Y_\iota}) \sum_{c|L_i,IV_i}^{n_c} \frac{X_{ij|c}}{n_c}$$

$$= \sum_j^{n_i} (Y_{ij} - \overline{Y_\iota}) (X_{ij} - \sum_{c|L_i,IV_i}^{n_c} \frac{X_{ij|c}}{n_c}) = \sum_j^{n_i} (Y_{ij} - \overline{Y_\iota}) (X_{ij} - E(X_{ij}|L_i, IV_i)).$$

Then, the overall statistic across families and its variance are,

$$T_W = \sum_i^{n_{\text{fam}}} T_i - E(T_i) = \sum_{i=1}^{n_{\text{fam}}} \sum_j^{n_i} (Y_{ij} - \overline{Y_\iota})(X_{ij} - E(X_{ij}|L_i, IV_i)) \quad (1)$$

$$\text{Var}(T_W) = \sum_i^{n_{\text{fam}}} \sum_j^{n_i} \sum_{j'}^{n_i} (Y_{ij} - \overline{Y_\iota})(Y_{ij'} - \overline{Y_\iota}) \text{Cov}(X_{ij}, X_{ij'}|L_i, IV_i))$$

where $\text{Cov}(X_{ij}, X_{ij'}|L_i, IV_i) = \frac{\sum_{c|L_i,IV_i}^{n_c} (X_{ij|c} - E(X_{ij}|L_i, IV_i))(X_{ij'|c} - E(X_{ij'}|L_i, IV_i))}{n_c}$. Under the null hypothesis that every founder chromosome is equally likely to be a carrier chromosome, i.e. that no variants are associated with the trait, TRACE_W is a score test and follows the standard normal distribution $\text{TRACE\_W} = \frac{T_W}{\sqrt{\text{Var}(T_W)}} \sim N(0,1)$

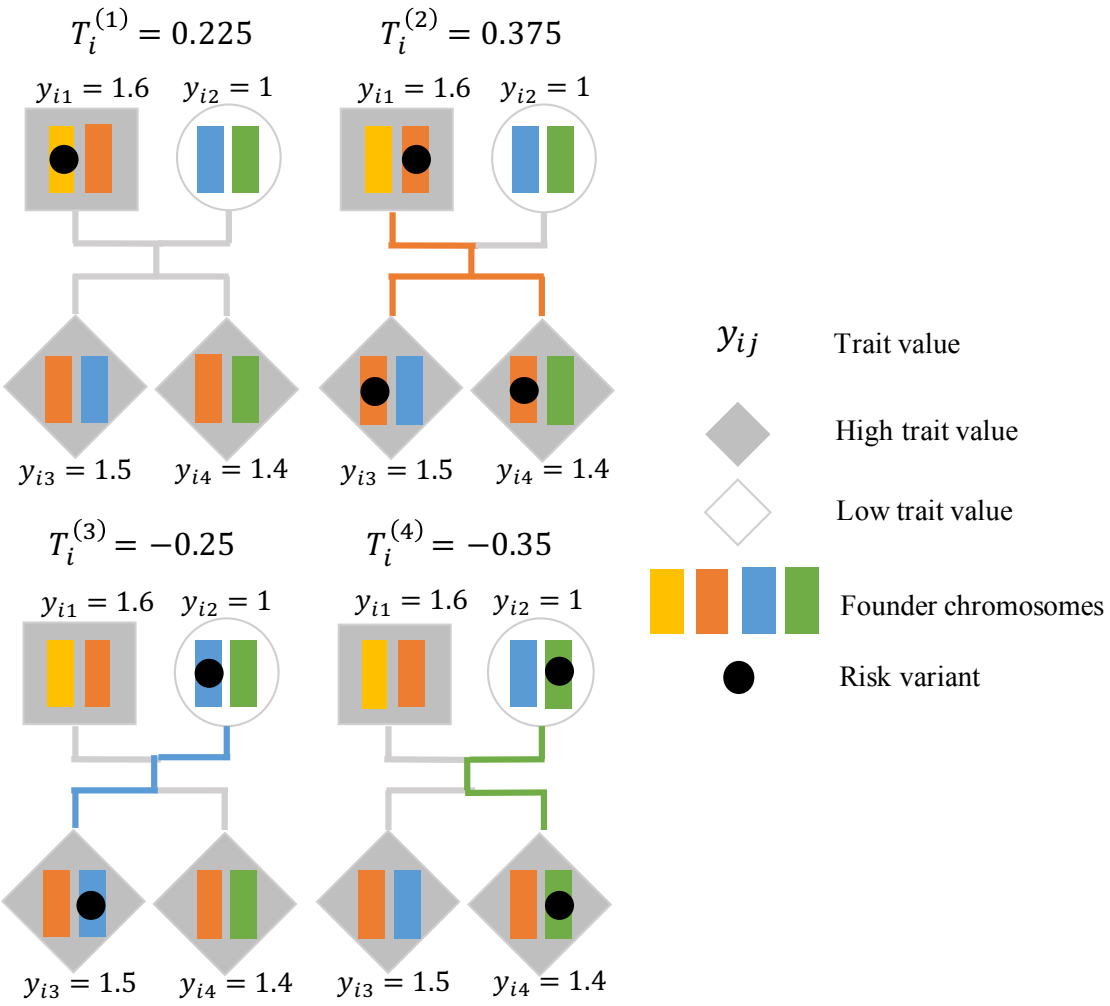


Figure 4-1. Illustration of inheritance vector for a family with two founders and two offspring assuming a single variant in the region. There are four possible paths to place the variant in founder chromosomes. The colored-line indicates the path that a variant is transmitted to offspring. Given that the variant is transmitted through a particular path k, $T_i^{(k)}$ is the summary statistic measuring the covariance between the trait and carrier chromosomes. Conditional on the inheritance vector in family i, we can thus enumerate all four possible $T_i^{(k)}$. Under the null hypothesis that each placement is equally likely to occur, we then can calculate $\mu_i = \frac{1}{4}(0.225 + 0.375 - 0.25 - 0.35) = 0$. Under the alternative that a variant increases the trait value, the variant is more likely to appear on the second (orange) founder chromosome and reach two offspring with high traits.

4.2.2. Include between-family information to form a combined test: TRACE

As shown in equation (1), within-family information calculates the covariance between the trait and carrier chromosomes $(Y_{ij} - \overline{Y_i})(X_{ij} - E(X_{ij}|L_i, IV_i)$. Define $\gamma_i = X_{ij} - E(X_{ij}|L_i, IV_i)$ and $\delta_i = E(X_{ij}|L_i, IV_i)$. δ_i is the family loading of carrier

chromosomes and equals the mean number of carrier chromosomes in a founder of the i^{th} family. Since accounting for the expected number of the transmitted carrier chromosomes in a family, within-family information γ_i represents the additionally transmitted carrier chromosomes and is independent of δ_i. Thus, we propose to use δ_i to test for the association between families. To test for the association using between-family information, we evaluate if the mean trait of all family members is associated with the family loading, $T_B = \sum_i^{n_{\text{fam}}}\left[(\overline{Y_i} - \bar{Y})(\delta_i - \bar{\delta})\right]$ where $\bar{\delta} = \frac{\sum_{i=1}^{n_{\text{fam}}} \delta_i}{n_{\text{fam}}}$ representing the mean family loading in the samples. And $\text{var}(T_B) = \sum_i^{n_{\text{fam}}}(\overline{Y_i} - \bar{Y})^2 \frac{2\hat{f}(1-\hat{f})}{n_i^{\text{founder}}}$, where $\hat{f}$ is the estimated carrier chromosome frequency based on all founders and n_i^{founder} is the number of founders in i^{th} family. Finally, we combine T_W and T_B to jointly test for the association,

$$\text{TRACE} = \frac{T_W + T_B}{\sqrt{\text{var}(T_W) + \text{var}(T_B)}} \sim N(0,1)$$

4.2.3. Simulation model

We use simulations to compare power between the proposed tests and existing family-based methods. We consider two existing family gene-based tests: Pedgene (Schaid et al. 2013), a gene-based test which implicitly uses both between- and within-family information by adjusting for the relatedness in families, and FB-SKAT that extends the within-family-approach FBAT to a gene-based test (Ionita-Laza et al. 2013). We consider two scenarios of sampling. First, families are sampled without ascertainment, i.e. we use every simulated family. Second, in the ascertained scenario, we only use families that have a pre-specified number of family members ascertained to have a trait in the top 10% percentile in the population.

Let y_{ij} denote the trait for the j^{th} member in the i^{th} family, $\underline{G_{ij}} = \{g_{ij1}, g_{ij2}, \dots, g_{ijm} | g_{ij\cdot} \in (0,1,2)\}$ an indicator vector for carrying the variant allele at m^{th} variant of interest. To generate families, we model the heritability in a family by incorporating the kinship matrix of i^{th} family $\boldsymbol{K}_i$ and use the trait model $y_{ij} = \alpha + \underline{G_{ij}}\underline{\beta} + F_{ij} + \varepsilon_{ij}$, where F_{ij} is the corresponding entry for j^{th} member in $2\boldsymbol{K}_i$, α is the mean, and $\underline{\beta} = (\beta_1, \beta_2, \dots, \beta_k)$ is the effect size of the risk variant assuming the effect size is a function of minor allele frequency (*MAF*): $r\frac{1}{4}\log_{10}(MAF)$; thus, a risk variant of *MAF* 0.0001 has effect size *r.* For neutral variants, the effect size is set as zero. In this disease model, the covariance matrix for i^{th} family is $\boldsymbol{V}_i = 2\sigma_{PE}^2\boldsymbol{K_i} + \sigma_E^2\boldsymbol{I}$, where $\sigma_{PE}^2 = 0.5$ is the shared polygenic and environmental effect in the family, and $\sigma_E^2 = 0.5$ is the non-measurable error.

We simulate sequence data using a coalescent-model-based simulator COSI (Schaffner et al. 2005) and generate ten thousand 1kb chromosomes. We randomly select causal variants where the summed allele frequency is 0.01 and fix these causal variants in the simulations. Assuming no recombination within a family, to generate $\underline{G_{ij}}$, we sample two chromosomes for each founder from those ten thousand chromosomes; then each offspring inherits two chromosomes following Mendel law (MacCluer et al. 1986). For non-ascertained scenarios, we simulate the phenotype for all members according to the trait model introduced above and keep every family until a specified number of families have been generated. For ascertained scenarios, we simulate the phenotype for all members and only keep the families that have a specified number of family members who have traits above the 90$^{\text{th}}$ percentile. We simulate family by family until we have generated a specified number of families. Then, to compare power, we apply the proposed tests and

the existing methods with different family structures. In the following sections, we use 1,000,000 replications to evaluate type I error rate and 1,000 replications for each scenario to calculate power.

4.3. Results

Using simulations generating 1kb regions, we consider two scenarios. First, families are collected without ascertainment. Second, families are generated with ascertainment where we keep only the family that has the specified number of members belonging in the top 10^{th} percentile. To compare power, we use different family structures from two-generation to three-generation pedigrees. First, we evaluate nominal type I error rate for all considered methods. Second, to quantify the power gain by including between-family information, we compare TRACE_W, which only considers within-family information, to TRACE, which considers both within- and between-family information. Third, we compare TRACE to existing methods, Pedgene and FBSKAT.

4.3.1. Type I error rate

We considered three pedigree structures in both ascertained and non-ascertained scenarios (see Methods). For two-generation pedigree, there were 4 members; for three-generation pedigree we considered 7 and 22 members. We coded each pedigree structure by the number of generations (g), number of family members (n), number of members with a trait above the 90^{th} percentile (a), and the number of members with a trait below the 90^{th} percentile (u). For example, under non-ascertained scenarios, a two-generation family with 4 members was denoted as 2g.4n; under ascertained scenarios, when there

were two members with a trait above the 90th percentile and two members below the 90th percentile, this pedigree was denoted as 2g.2a.2u. Moreover, any family member can have the trait above the 90th percentile. The pedigree structures can be found in Appendix C.1. To ease the computational burden, type I error rate was set at 2.5×10^{-4}. To compare results across pedigrees, given sequenced founders in a family, we assume that we can accurately impute offspring genotypes using genotyping array with a negligible cost relative to sequencing; we thus considered a fixed number of 2000 founders sequenced for each pedigree structure, resulting in 1000, 667, and 286 families for 2g.4n (2g.2a.2u), 3g.7n (3g.3a.4u), and 3g.22n (3g.5a.17u), respectively. FB-SKAT failed to run for the large pedigree 3g.5a.17u, and thus was excluded for the result of large pedigree. Across the pedigree structures, all considered methods maintained nominal type I error as shown in Table 4-1.

	Non-ascertained			Ascertained		
	2g.4n	3g.7n.4n	3g.22n	2g.2a.2u	3g.3a.4u	3g.5a.17u
TRACE_W	1.0×10^{-4}	1.2×10^{-4}	1.5×10^{-4}	1.0×10^{-4}	1.1×10^{-4}	1.3×10^{-4}
TRACE	1.0×10^{-4}	1.2×10^{-4}	1.4×10^{-4}	1.1×10^{-4}	1.1×10^{-4}	1.4×10^{-4}
Pedgene	1.6×10^{-4}	1.5×10^{-4}	1.4×10^{-4}	1.7×10^{-4}	1.6×10^{-4}	1.4×10^{-4}
FB-SKAT	1.1×10^{-4}	1.4×10^{-4}	1.3×10^{-4}	1.1×10^{-4}	1.4×10^{-4}	1.2×10^{-4}

Table 4-1. Type I error rate evaluation for TRACE_W, TRACE, Pedgene, and FB-SKAT for carrier chromosome frequency 0.01 under different family structures and nominal $\alpha = 2.5\times10^{-4}$.

4.3.2. Power gain by including between-family information

Since between- and within-family information provide independent evidence about the association, including between-family information can increase power compared to only using within-family information. We simulated families for each pedigree structures described above and compared TRACE_W to TRACE. As shown in Figure 4-2, we found that the power gain by including between-family information was more substantial for

non-ascertained than ascertained scenarios. For example, for two-generation pedigree, the maximum difference in power between TRACE_W to TRACE was 0.45 under no ascertainment and decreased to 0.08 when two members were ascertained in > 90th percentile. The power gain also depended on the pedigree structure. When comparing medium-size pedigree 3g.7n to small pedigree 2g.4n, the power gain was more substantial for small pedigree than medium pedigree; the maximum difference between TRACE_W to TRACE was 0.45 for 2g.4n compared to 0.15 for 3g.7n. For large pedigree 3g.22n, there was little power gain by including between-family information for both non-ascertained and ascertained scenarios.

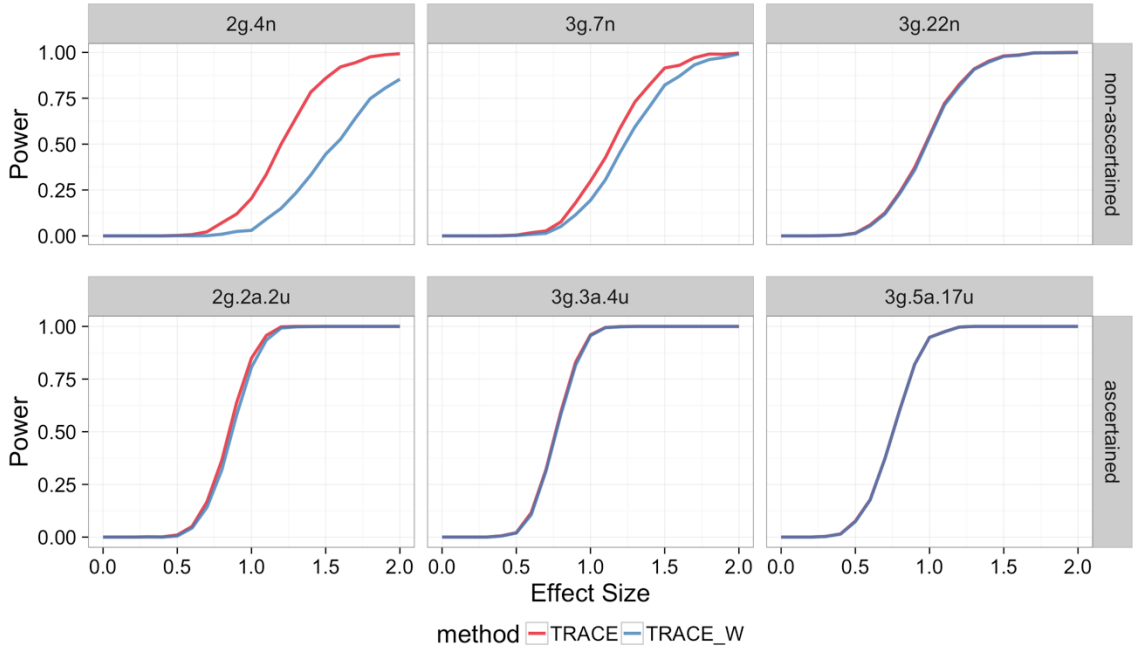


Figure 4-2 The power comparison evaluated at $\alpha = 2.5\times10^{-6}$ between the within-family test TRACE_W and the combined test TRACE under non-ascertained and ascertained scenarios. Power curve is shown as a function of effect size of risk variants in a gene with carrier chromosome frequency 0.01.

4.3.3. Power comparison across TRACE, Pedgene, and FBSKAT

Under different pedigree structures and ascertainment, we compared TRACE to

existing family-based methods Pedgene and FB-SKAT; we excluded FB-SKAT for large pedigree because it failed to complete the analyses. Throughout the considered scenarios in Figure 4-3, FB-SKAT was the least powerful method. For Pedgene and TRACE, the power difference depended on the size of the pedigree structure and ascertainment. Under the non-ascertained scenario, Pedgene was the most powerful method for 2g.4n and 3g.7n followed by TRACE. For example, given effect size 1.3 and pedigree 2g.4n, the power for Pedgene and TRACE was 0.80 and 0.65, respectively. In contrast, for large pedigree 3g.22n, TRACE was more powerful than Pedgene; for effect size of 1.0, the power for TRACE and Pedgene was 0.56 and 0.51. In this case, within-family contained the most information, specifically the sharing of rare variants between family members. TRACE could directly use this within-family information through inheritance vector to up-weight the sharing of rare variants and outperform Pedgene, which does not exploit this information.

For ascertained scenarios, the power of TRACE substantially increased and it was consistently the most powerful method from small to large pedigree settings. For example, for 2g.2a.2u and effect size 0.8, the power for TRACE and Pedgene was 0.64 and 0.59. The advantage in power for TRAP was more substantial for large pedigree 3g.5a.17u than small pedigree 2g.2a.2u. The maximum power difference between TRACE and Pedgene was 0.12 for 3g.5a.17u compared to 0.06 for 2g.2a.2u.

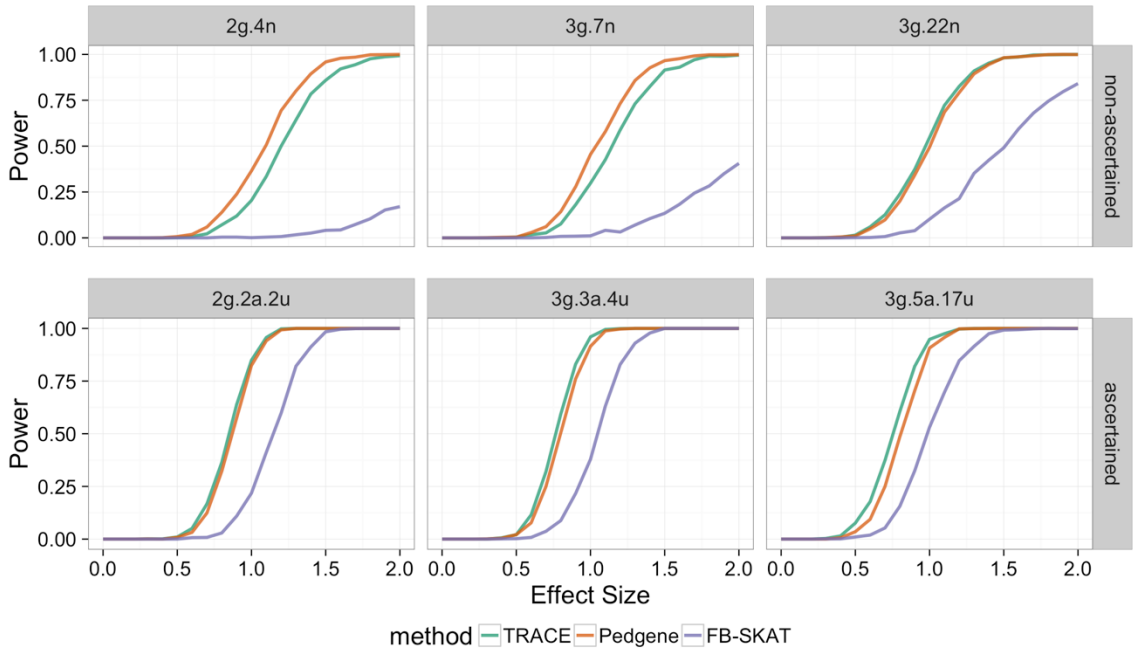


Figure 4-3 The power comparison evaluated at $\alpha = 2.5\times10^{-6}$ across TRAP, Pedgene and FB-SKAT under non-ascertained and ascertained scenarios. Power curve is shown as a function of effect size of risk variants in a gene with carrier chromosome frequency f = 0.01.

4.4. Discussion

In this chapter, we developed a novel family-based rare-variant association test for continuous traits that can flexibly be applied to arbitrary pedigree structures and efficiently use within-family information. We use the fact that for family members with a similar trait value, they are more likely to share the associated variant. Conditional on the transmission vector, we evaluate if the sharing of the variant is associated with the trait similarity. Intuitively, if a variant is not associated with the disease, it would be equally likely transmitted to the offspring through every founder chromosome; thus, we evaluate if the variant is biasedly transmitted to the offspring through a specific founder chromosome. Moreover, TRACE combined within- and between-family information to fully use family information to jointly test for associations.

TRACE is a combined test that use both within- and between-family information. We have showed that TRACE can be useful in both non-ascertained and, particularly ascertained scenarios. For the ascertained scenario, within-family information provides the most evidence about association; thus, there was a minor power gain as comparing TRACE to TRACE_W, which only used within-family information. TRACE was more powerful than existing family-based methods by better using the within-family information through inheritance vector to exploit the sharing of rare variants between family members. Under the ascertained scenario, family members with a high-level trait were more likely to share a variant that increases the trait value, leading to power gain. In addition, due to ascertainment, it was more likely that a family segregates at least one risk variant, leading to more informative families in the test, particularly for large pedigrees with multiple founders; thus, we observe a greater power gain for large pedigrees than small pedigrees. In general, Pedgene reduced the effective sample size to account for relatedness and did not consider the sharing of rare variants in a family; thus, the larger the pedigree, the greater the power difference was between TRACE and Pedgene. Although FB-SKAT should benefit from ascertainment, FB-SKAT was less powerful than TRACE as it does not efficiently use within-family information and breaks an extended family into trios. In addition, FB-SKAT cannot consider parents' phenotypic information.

Under non-ascertained scenarios, TRACE includes between-family information and substantially improves power over TRACE_W, especially for small pedigree structures. Although Pedgene reduced the effective sample size to account for relatedness, we only observed TRACE being more powerful than Pedgene for large pedigrees. For a large pedigree, sharing an IBD variant among many family members across generations

provides stronger evidence that the variant was associated with disease compared to small pedigree; thus, TRACE can better exploit within-family information to up-weight this sharing evidence as compared to Pedgene (Appendix C.2).

By including between-family information, TRACE can substantially increase power over TRACE_W, which only uses within-family information, particularly for small pedigrees. We note that including between-family information is more prone to population stratification due to the heterogeneity in origin between families than only using within-family information. However, TRACE can easily adapt the existing strategy to overcome the common confounders in association studies. For example, by taking the residual of the trait adjusted for principal components and available covariates as the new trait, TRACE can adapt this existing strategy to reduce false positive signals due to population stratification.

In conclusion, we propose a new powerful approach to better use family information for testing rare variant associations with continuous traits. With ascertained families, TRACE is more powerful than the existing family-based methods; TRACE is also useful for non-ascertained families with multiple members and generations. As discussed in Chapter 3, TRACE relies on inheritance vectors to evaluate associations and inferring inheritance vectors for large pedigrees can be challenging; thus, future work is warranted to investigate and accommodate TRACE to scale up to future large family-based association studies.

Chapter 5
Discussion

Rare variants have been hypothesized to explain part of the missing heritability of complex diseases (Manolio et al. 2009). With the innovation of sequencing technology, recent studies have routinely assessed rare variants using next-generation sequencing technology (Sidore et al. 2015; Fuchsberger et al. 2016; Fritsche et al. 2016). However, identifying associated rare variants faces the challenge of low statistical power unless using a very large sample sizes (Lee et al. 2014). In this dissertation, we have presented a series of methods to harness the disproportionate sharing of rare variants between family members to increase power to detect rare variant associations. In this chapter, we discuss the significance of each method, their extensions and the impact on future large-scale family-based studies.

In Chapter 2, we presented a new paradigm TRAFIC to use identity-by-descent (IBD) information in affected sibpairs to test for rare variant associations. For a rare risk variant, it is expected that only one parent in a family carries the variant; thus, two affected siblings share the same variant, i.e. the rare variant resides on the IBD chromosome. Using this fact, we test whether rare variants are equally likely to be observed on IBD chromosomes and non-IBD chromosomes in affected sibpairs, and found TRAFIC can be substantially more powerful in detecting rare variants than using population case-control

samples. Furthermore, as shown in Zawistowski et al. (2014) and our results, minor population stratification can substantially inflate type I error rate and thereby generate spurious association signals. In rare variant association studies, adjusting for principal components (Price et al. 2006) may not be sufficient to control for population stratification (Mathieson and McVean 2012; Liu, Nicolae, and Chen 2013). An alternative way to prevent population stratification is to use prudently matched cases and controls. Intuitively, TRAFIC uses non-IBD chromosomes to serve as cautiously matched "controls" compared to IBD-chromosomes as "cases" in each sibpair to be robust to population stratification.

An innovation of TRAFIC is that we can translate an affected-sibpair design into a "case-control" design by assigning a new label to IBD chromosomes as "cases" and non-IBD chromosomes as "controls." This connection allows an affected-sibpair design to take advantage of existing population gene-based methods (Wu et al. 2011; Zawistowski et al. 2010; Price et al. 2010; Lee et al. 2014), for example, to incorporate the variants with different directions of effect. Furthermore, this translation into a "case-control" design also enables the inclusion of external population samples, generated by many existing and ongoing sequencing projects, to further increase power. Recently, by matching with the inferred ancestry, using external population controls have been proposed to increase sample sizes (C. Wang et al. 2014; Bodea et al. 2016). Although future evaluations of the matching algorithms to avoid population stratification for rare variant associations are required, in principle, given prudently matched external controls, TRAFIC can further improve power to identify associated rare variants by comparing IBD-chromosomes as "cases" to the external controls.

To fully harness the pedigree information contained in extended families, in Chapter 3

and Chapter 4, we devised TRAP and TRACE to efficiently use pedigree information for binary and continuous traits, respectively. The innovation of TRAP and TRACE is to better use within-family information through inheritance vectors, which indicate transmission paths by which a variant can be transmitted from founders to offspring. If a variant increases the disease risk (trait value), the variant would be disproportionally transmitted through the path that can include the most family members with the disease (a high trait value). In a scenario that multiple members in a family are ascertained to have the disease (a high trait value), these ascertained family members tend to share the associated rare variants. Because TRAP and TRACE can efficiently use the inheritance vector to up-weight the sharing of rare variants between family members, we showed that TRAP and TRACE can be more powerful than existing family-based methods from small to large pedigree structures. In addition, as the sharing of rare variants across generations in a family is strong evidence for associations, the power gain increases with the size of pedigree. For a large pedigree with multiple generations, even without ascertainment, we found that TRACE is more powerful than existing family-based methods which do not directly take advantage of the sharing of variants among family members.

Recent work has suggested that rare variants play an critical role in contributing to the variance of traits such as height (Yang et al. 2015) and prostate cancer (Mancuso et al. 2016). At the same time, it has been shown that the exponential growth of human population and selection can alter the genetic architecture of a trait in which rare variants, such as singletons, account for the majority of the trait variance (Uricchio et al. 2016). In such a scenario, existing population gene-based tests have inadequate statistical performance (Uricchio et al. 2016). Alternatively, to achieve sufficient power to detect the

associations, a family-based design might be necessary to increase observed copies of rare variants in families and then employ methods, such as TRAP and TRACE, to efficiently use within-family information to leverage the sharing of rare variants, leading to the most power gain.

Although the methods presented in this dissertation use IBD information to improve power for rare variant associations, we note that generated results are independent from how linkage studies use IBD information, specifically allele-sharing linkage methods. In principle, allele-sharing linkage methods evaluate whether there is an excessive number of IBD chromosome shared by affected family members (Weeks and Lange 1988; Whittemore and Halpern 1994; Sham et al. 2002). In contrast, TRAFIC, TRAP, and TRACE are conditional on the excessive number of shared IBD chromosomes, and then test for whether IBD chromosomes that are shared by multiple affected family members are more likely to carry rare variants. This implies the association signal is driven by the linkage-disequilibrium between the tested variants and the causal variant as used in association studies. Thus, as in association studies, TRAFIC, TRAP, and TRACE provide a finer resolution than the resolution in linkage studies (Laird and Lange 2006).

With the growing volume of genetic studies that collect tens of thousands samples (Lee et al. 2014; R. Chen et al. 2016), the success of future family-based studies depends on the scalability of family-based methods applied to a large sample of families. Linear mixed models or logistic mixed models are popular approaches to account for the relatedness in family-based studies, and many methods have been developed based on the mixed model approach (Kang et al. 2010; Zhou and Stephens 2012; Eu-ahsunthornwattana et al. 2014; H. Chen et al. 2016). Fitting these mixed models requires

one step to calculate the likelihood under the null hypothesis, which has a computational time complexity that is cubic to the sample size (Eu-ahsunthornwattana et al. 2014; H. Chen et al. 2016), and can be computationally difficult for a very large number of samples (>100,000). On the other hand, when using within-family information, TRAP and TRACE only examine those families which segregate a rare variant, and thus these methods need much fewer computational resources.

However, TRAP and TRACE require inheritance vectors as input. To infer inheritance vectors, the computational burden of existing methods grows exponentially with the family size (number of non-founders) or the number of variants considered (X. Li and Li 2011; Ott, Wang, and Leal 2015). For example, for ten families in a 1kb region with 50 variants, using Merlin (Abecasis et al. 2002) to infer two-generation pedigree consisting of 2 founders and 2 offspring takes <0.001 second; in contrast, the time to infer inheritance vectors for two-generation pedigree with 3 founders and 4 offspring, and three-generation pedigree of 7 founders and 15 offspring are 0.01 and 108 seconds, respectively. Although new methods have been proposed to reduce the computational burden to be linear with the number of non-founders (X. Li and Li 2011; O'Connell et al. 2014), it can still impose a computational burden in a sample with many extended families. In addition, the inference of inheritance vectors for large pedigrees can be challenging when there are multiple missing founders across generations. To account for the missingness, enumerating all possible inheritance patterns among missing founders in several generations quickly becomes computationally prohibitive. Besides, the presence of missing founders elevates the chance of falsely reporting the sharing of chromosomes between offspring, leading to power loss. Thus, future work is warranted

to investigate the scalability of the necessary steps, and accommodate TRAP and TRACE to apply in such scenarios.

In conclusion, with the abundant number of rare variants discovered by existing and ongoing sequencing projects, drawing further findings on rare variants contributing to complex diseases requires robust and powerful methods for association studies. In this dissertation, we have presented three statistical methods, each aimed at using the sharing of rare variants in family data to advance our understanding of etiologies and promote new preventive and therapeutic strategies for complex diseases.

Appendix A
For Chapter 2

A.1. EM algorithm and multiple imputation

This imputation estimates the minor allele frequencies on shared chromosome regions p_S and non-shared chromosome regions p_{NS} at a single SNP position. Let N_S be the number of shared chromosome regions, x_S be the number of minor alleles located on IBD chromosome regions, N_{NS} be the number of non-IBD chromosome regions, and x_{NS} be the number of minor alleles located on non-IBD chromosome regions. Then x_S, x_{NS} are binomially distributed with the full likelihood function,

$$L(x_S, x_{NS}|N_S, N_{NS}, p_S, p_{NS}) = \binom{N_S}{x_S} p_S^{x_S}(1-p_S)^{N_S-x_S}\binom{N_{NS}}{x_{NS}} p_{NS}^{x_{NS}}(1-p_{NS})^{N_{NS}-x_{NS}}.$$

Let k_S be the known total number of shared alleles on IBD=2 sibpairs (sibpairs who share two IBD chromosome regions), and k_{NS} be the known total number of non-shared alleles on IBD=0 and single allele on IBD=1 sibpairs. Suppose there are u double-heterozygote IBD=1 sibpairs. Among those sibpairs, u_S number of sibpairs share an allele and u_{NS} of them do not share an allele. The likelihood function of observing the total number of shared allele x_S and non-shared allele x_{NS} can be rewritten as

$$\binom{N_S}{k_S+u_S} p_S^{k_S+u_S}(1-p_S)^{N_S-k_S-u_S}\binom{N_{NS}}{k_{NS}+2u_{NS}} p_{NS}^{k_{NS}+2u_{NS}}(1-p_{NS})^{N_{NS}-k_{NS}-2u_{NS}}.$$

To set up the initial values for the algorithm, we estimate the allele frequency on IBD

chromosome regions p_S by counting the total number of non-shared alleles divided by the number of IBD=2 sibpairs. Similarly, we estimate the allele frequency on non-IBD chromosome regions p_{NS} from IBD=0 sibpairs.

In the E-step, we calculate the expectation of double-heterozygote IBD=1 sibpairs who share and don't shared an allele conditional on the current estimation of parameters,

$$E(u_S|p_S, p_{NS}) = u\frac{p_S(1-p_{NS})^2}{p_S(1-p_{NS})^2 + (1-p_s)p_{NS}^2}$$
$$E(u_{NS}|p_S, p_{NS}) = u\frac{(1-p_s)p_{NS}^2}{p_S(1-p_{NS})^2 + (1-p_s)p_{NS}^2}.$$

In the M-step, we update new p_S and p_{NS} as the solution of maximum likelihood estimators of the expected likelihood function

$$p_S^{new} = \frac{\mathrm{k}_S + u_S}{N_S}, \quad p_{NS}^{new} = \frac{k_{NS} + 2u_{NS}}{N_{NS}}.$$

The algorithm repeats between E-step and M-step until the estimation of parameters converges. To impute the allele sharing status for double-heterozygote IBD=1 sibpairs, let p^* the probability of being a shared allele given a double-heterozygote IBD=1 sibpair,

$$p^* = \frac{p_S(1-p_{NS})^2}{p_S(1-p_{NS})^2 + (1-p_s)p_{NS}^2}.$$

The procedures of multiple imputation are as follows:

1. Draw the number of double-heterozygote IBD=1 sibpairs who share an allele μ_s from Binomial(μ, p^*)

   Then the number of double-heterozygote IBD=1 sibpairs who do not share an allele is

   $\mu - \mu_s$
2. With the imputed-complete data, calculate the test statistic T_i and its variance σ_i^2
3. Repeat the above procedures for *D*=10 times
4. Combine *D* test statistics and their variances by the following rule

$$T_D = \frac{\sum_{i=1}^{D} T_i}{D}$$

$$\sigma_D^2 = \frac{1}{D}\sum_{i=1}^{D}\sigma_i^2 + (1 + 1/D)\frac{1}{D-1}\sum_{i=1}^{D}(T_i - T_D)^2$$

5. Use T_D and σ_D^2 to perform the hypothesis.

A.2. Simulation result for Imputing Double Heterozygotes

We evaluated multiple imputation assuming one single underlying risk variant with minor allele frequency *f* = 0.01. Conditional on *f*, a single variant represents the most challenging scenario as, the probability of a double-heterozygote at a single position in sibpairs that share one chromosome region IBD decreases with the number of underlying variants in the region and the probability that such a double heterozygote is the result of two non-shared variants also decreases. In Table A-1, we compared three models: (1) the sharing status of alleles was known (true), (2) assuming all double-heterozygote sibpairs were the results of shared alleles (naive estimate), and (3) using the multiple imputed sharing status (corrected estimate). We set the false positive rate at 0.05. The test was well calibrated if sharing status was known. After applying multiple imputation, there was no inflation on the false positive rate. For example, under the null hypothesis of $\mu = 1 \text{ and } \sigma^2 = 0$, for a single rare variant (*f* = 0.01), the false positive rate for the true, naïve and corrected estimate were 0.058, 0.058, and 0.055, respectively. For more common variants (*f* = 0.2), the inflation (0.545) was substantial when using the naive estimate of sharing status but remained at the nominal level for the true model and the corrected estimate at 0.049.

Under the alternative with $\mu > 1$, the loss in power due to imputation compared to the true model was negligible for low *f*. For example, for a single rare risk variant (*f* = 0.01, $\mu = 1.5$), the power of the true, and corrected were 0.294, and 0.296, respectively. For a single common risk variant (*f* = 0.20, $\mu = 1.2$) the power was 0.557, and 0.467.

	True	Naive	Corrected		True	Naive	Corrected
***f*=0.01**				***f*=0.1**			
$\boldsymbol{\mu = 1}$	0.058	0.058	0.055	$\boldsymbol{\mu = 1}$	0.054	0.121	0.060
$\boldsymbol{\mu = 1.5}$	0.294	0.313	0.296	$\boldsymbol{\mu = 1.2}$	0.356	0.691	0.324
$\boldsymbol{\mu = 2}$	0.835	0.847	0.815	$\boldsymbol{\mu = 1.4}$	0.888	0.992	0.827
***f*=0.05**				***f*=0.2**			
$\boldsymbol{\mu = 1}$	0.047	0.057	0.046	$\boldsymbol{\mu = 1}$	0.049	0.545	0.049
$\boldsymbol{\mu = 1.3}$	0.464	0.602	0.433	$\boldsymbol{\mu = 1.1}$	0.205	0.903	0.169
$\boldsymbol{\mu = 1.5}$	0.847	0.929	0.806	$\boldsymbol{\mu = 1.2}$	0.557	0.993	0.467

Table A-1. The simulated false positive rate and power for TRAFIC using true model, naïve estimate and corrected multiple imputation under different summed allele frequencies f and mean relative risk μ. The true model assumes the sharing status of alleles is known. The naive estimate treats all ambiguous variants as shared, and the corrected multiple imputation uses EM followed by multiple imputation to perform the hypothesis testing. The sample size was 1000 sibpairs and false positive rate was set at 0.05.

A.3. Calculating $\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS} | \mathrm{AA}_R, \mathrm{S})$

The power of TRAFIC depends on

$$\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS} | \mathrm{AA}_R, \mathrm{S}) = \mathrm{P}(\mathrm{AA}_\mathrm{R} | \mathrm{H}_S, \mathrm{H}_{NS}) \mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS} | \mathrm{S}) \mathrm{P}(\mathrm{S}) \frac{1}{\mathrm{P}(\mathrm{AA}_R, \mathrm{S})}$$

where $\mathrm{P}(\mathrm{AA}_\mathrm{R} | \mathrm{H}_S, \mathrm{H}_{NS})$ depends on the underlying genetic and the effect size model described as below, and $\mathrm{P}(\mathrm{S})$ is the segregating probability: $\mathrm{P}(\mathrm{S} = 0) = 0.25$, $\mathrm{P}(\mathrm{S} = 1) = 0.50$ and $\mathrm{P}(\mathrm{S} = 2) = 0.25$. $\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS} | \mathrm{S})$ is described at the end of this section.

A.3.1 Genetic model

Suppose the population has the disease prevalence K. Let KK_R denote the recurrence risk for a sibpair. The penetrance contributed from the locus of interest is K_L and the

contribution to the recurrence risk is $K_L K_{LR}$. Similarly, the penetrance contributed from the rest of the genome is defined as K_G, and the contribution to the recurrence risk is $K_G K_{GR}$. Assume there are m distinct risk haplotypes $h_1 \ldots h_m$. Let $\omega(h_s, h_t)$ be the penetrance component of any genotype $h_s h_t$. Under the multiplicative model, each locus contributes independently to the heritability. The recurrence risk can be expressed as $KK_R = K_L K_{LR} K_G K_{GR}$. The probability of observing an affected sibpair given haplotypes $(h_i, h_j), (h_k, h_l)$ is

$$\mathrm{P}\left(\mathrm{AA_R} \middle| h_i, h_j, h_k, h_l\right) = \omega\left(h_i, h_j\right)\omega(h_k, h_l) K_G K_{GR}.$$

A.3.2 Effect size model

We model the effect size (relative risk) ω of haplotype as a random variable following a distribution $g(.)$ with the first two moments μ and σ^2 known. The penetrance for haplotype (h_i, h_j) is the product of both haplotypes' effect i.e. $\omega\left(h_i, h_j\right) = \omega_i \omega_j$. For those haplotypes with no risk alleles, the relative risk is set to 1. Then we can further express $\mathrm{P}(\mathrm{AA_R}|\mathrm{H}_S, \mathrm{H}_{NS}) = \omega_1\omega_2\omega_3\omega_4$ as a function of μ and σ^2 by using the underlying genetic and effect model.

A.3.3 Calculating $\mathbf{P(AA_R|H_S, H_{NS})}$

When considering the contribution from the non-shared chromosome regions, we take the average on the range of all possible effect sizes $\int \omega g(\omega) d\omega = \mu$. In a sibpair who shared an IBD chromosome region, the IBD chromosome region is observed twice. Thus, the contribution of these chromosome regions to the overall penetrance would involve the term $\int \omega_s^2 g(\omega_s) d\omega_s = \mu^2 + \sigma^2$. In the following, we calculate the penetrance given the haplotype under multiplicative model.

$$\mathrm{P}(\mathrm{AA_R}|\mathrm{H}_S = h_s, \mathrm{H_{NS}} = h_{ns},)$$

$$\propto \left[\int \omega_s^2 g(\omega_s) d\omega_s\right]^{h_s} \left[\int \omega g(\omega) d\omega\right]^{h_{ns}} = (\mu^2 + \sigma^2)^{h_s} \mu^{h_{ns}}$$

A.3.4 Calculating $\mathbf{P(H_S, H_{NS}, |S)}$

Given S, the frequency *f* of carrying at least one allele (risk haplotype) is identical on the shared and non-shared chromosome regions,

$$\mathrm{P}(\mathrm{H}_S = h_s, \mathrm{H_{NS}} = h_n|\mathrm{S} = 0) = f^{h_n}(1-f)^{4-h_n}$$

$$\mathrm{P}(\mathrm{H}_S = h_s, \mathrm{H_{NS}} = h_n|\mathrm{S} = 1) = f^{h_s+h_{ns}}(1-f)^{3-h_s-h_{ns}}$$

$$\mathrm{P}(\mathrm{H}_S = h_s, \mathrm{H_{NS}} = h_n|\mathrm{S} = 1) = f^{h_s}(1-f)^{2-h_s}$$

A.4. sMAF and analytical power calculation

In this study, we use a simple Chi-squared test to determine if the proportion of risk haplotypes is different on shared and non-shared chromosome regions. The power of the test depends on the non-centrality parameter λ, where

$$\lambda = \left(\frac{-(p_2 - p_1 - p0) \times \left(N \times \frac{N_S}{N} \times \frac{N_{NS}}{N}\right)^{\frac{1}{2}}}{\left(\frac{N_{NS}}{N} \times p_1 \times (1-p_1) + \frac{N_S}{N} \times p_2 \times (1-p_2)\right)^{\frac{1}{2}}} \right)^2$$

N_S and N_{NS} are the number of shared (cases) and non-shared (controls) chromosome regions, respectively. *N* denotes the sum of independent chromosome regions which equals $N_S + N_{NS}$. p_1 and p_2 are the proportions of shared IBD chromosome regions and non-shared IBD chromosome regions carrying a risk haplotype, respectively. The expected number of independent chromosome regions depends on $\mathrm{P}(\mathrm{S}|\mathrm{AA_R})$ which can be derived by integrating out H_S and H_{NS} in $\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}, S|\mathrm{AA_R}) =$

$\mathrm{P}(\mathrm{AA_R}|\mathrm{H}_S, \mathrm{H}_{NS})\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}|\mathrm{S})\mathrm{P}(\mathrm{S})\frac{1}{\mathrm{P}(\mathrm{AA_R})}$.

Suppose there are N pairs of affected siblings, the expected number of shared chromosome regions N_S (cases) is $N\times\mathrm{P}(\mathrm{S}=1|\mathrm{AA_R}) + 2N\times\mathrm{P}(\mathrm{S}=2|\mathrm{AA_R})$ and the expectation of non-shared IBD chromosome regions N_{NS} (controls) is $2N\times \mathrm{P}(\mathrm{S}=1|\mathrm{AA_R}) + 4N\times\mathrm{P}(\mathrm{S}=0|\mathrm{AA_R})$. The expectations of total number of shared and non-shared chromosome regions carrying a risk haplotype are $E[H_S] = N\times\mathrm{E}(\mathrm{H}_S|\mathrm{AA_R}) = \mathrm{N}\times[\mathrm{P}(\mathrm{H}_S=1|\mathrm{AA_R}) + 2\times\mathrm{P}(\mathrm{H}_S=1|\mathrm{AA_R})]$ and $E[H_{NS}] = N\times[\mathrm{P}(\mathrm{H}_{NS}=1|\mathrm{AA_R}) + 2\mathrm{P}(\mathrm{H}_{NS}=2|\mathrm{AA_R}) + 3\mathrm{P}(\mathrm{H}_{NS}=3|\mathrm{AA_R}) + 4\mathrm{P}(\mathrm{H}_{NS}=4|\mathrm{AA_R})]$ respectively; the marginal probability of H_S and H_{NS} can be derived by integrating out the other variable in $\mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}|\mathrm{AA_R}) = \sum_S \mathrm{P}(\mathrm{H}_S, \mathrm{H}_{NS}, S|\mathrm{AA_R})$. Thus, sMAF for cases and controls is $E[H_S]/\mathrm{E}[N_S]$ and $E[H_{NS}]/E[N_{NS}]$, respectively.

$$E[H_S]/\mathrm{E}[N_S] = \frac{\mathrm{P}(\mathrm{H}_S=1|\mathrm{AA_R})+2\times\mathrm{P}(\mathrm{H}_S=1|\mathrm{AA_R})}{\mathrm{P}(\mathrm{S}=1|\mathrm{AA_R})+2\times\mathrm{P}(\mathrm{S}=2|\mathrm{AA_R})}$$

where

$$\mathrm{P}(\mathrm{H}_S=1|\mathrm{AA_R}) = (\mu^2+\sigma^2)*\left((1-f)+\mu f\right)^2 * f * 0.5 + (\mu^2+\sigma^2)*2f(1-f) * 0.25$$

$$\mathrm{P}(\mathrm{H}_S=2|\mathrm{AA_R}) = (\mu^2+\sigma^2)^2 * f^2 * 0.25$$

$$\mathrm{P}(\mathrm{S}=1|\mathrm{AA_R}) = 0.5*\left(1+f*(\mu-1)\right)^2 * \left(1+f*(\mu^2+\sigma^2-1)\right)$$

$$\mathrm{P}(\mathrm{S}=2|\mathrm{AA_R}) = 0.25*\left(1+f*(\mu^2+\sigma^2-1)\right)^2$$

$$E[H_{NS}]/E[N_{NS}] = \frac{\mathrm{P}(\mathrm{H}_{NS}=1|\mathrm{AA_R})+2\mathrm{P}(\mathrm{H}_{NS}=2|\mathrm{AA_R})+3\mathrm{P}(\mathrm{H}_{NS}=3|\mathrm{AA_R})+4\mathrm{P}(\mathrm{H}_{NS}=4|\mathrm{AA_R})}{2\mathrm{P}(\mathrm{S}=1|\mathrm{AA_R})+4\mathrm{P}(\mathrm{S}=0|\mathrm{AA_R})},$$

where

$$\mathrm{P}(\mathrm{H}_{NS} = 1|\mathrm{AA}_{\mathrm{R}})$$

$$= (\mu * 4 * f * (1-f)^3 * 0.25 \; + \; (\mu * (1-f) + \mu * (\mu^2 \; + \sigma^2) * f)$$

$$* \, 2f(1-f) * 0.5)$$

$$\mathrm{P}(\mathrm{H}_{NS} = 2|\mathrm{AA}_{\mathrm{R}})$$

$$= (\mu^2 * 6f^2 * (1-f)^2 * 0.25 \; + (\mu^2 * (1-f) + \mu^2 * (\mu^2 + \sigma^2) * f)$$

$$* \, f^2 * 0.5)$$

$$\mathrm{P}(\mathrm{H}_{NS} = 3|\mathrm{AA}_{\mathrm{R}}) = \mu^3 \; * 4f^3(1-f) * 0.25$$

$$\mathrm{P}(\mathrm{H}_{NS} = 4|\mathrm{AA}_{\mathrm{R}}) = \mu^4 * f^4 * 0.25$$

$$\mathrm{P}(\mathrm{S} = 1|\mathrm{AA}_{\mathrm{R}}) = 0.5 * \big(1 + f * (\mu - 1)\big)^2 * \big(1 + f * (\mu^2 + \sigma^2 - 1)\big)$$

$$\mathrm{P}(\mathrm{S} = 0|\mathrm{AA}_{\mathrm{R}}) = 0.25 * \big(1 + f * (\mu - 1)\big)^4$$

The test statistic Z is,

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{N_s} + \frac{1}{N_{NS}})}}, \quad \text{where} \quad p = \frac{p_1 * N_s + p_2 * N_{Ns}}{N_s + N_{Ns}}, p_1$$

$$= E[H_S]/\mathrm{E}[N_S] \text{ and } p_2 = E[H_{NS}]/E[N_{NS}]$$

and Z^2 follows a Chi-square distribution with degree of freedom 1.

A.5. Gene-gene interaction

In a two-locus model, the joint effect from locus of interest L and remaining genome G is defined as

$$P(A|h_m, h_n, g_s, g_t) \propto \beta_L^{h_m + h_n} \, \beta_G^{g_s + g_t} \gamma^{(h_m + h_n)(g_s + g_t)}$$

Under this model, the marginal relative risk at locus L is

$$\frac{P(A|h_m=1)}{P(A|h_m=0)}=\frac{\beta_L\sum_{h_n}\beta_L^{h_n}p(h_n)\sum_{g_s}\sum_{g_t}(\beta_G\gamma)^{g_s+g_t}\gamma^{h_n(g_s+g_t)}p(g_s)p(g_t)}{\sum_{h_n}\beta_L^{h_n}p(h_n)\sum_{g_s}\sum_{g_t}\beta_G{}^{g_s+g_t}\gamma^{h_n(g_s+g_t)}p(g_s)p(g_t)},$$

and the marginal relative risk at locus G is

$$\frac{P(A|g_s=1)}{P(A|g_s=0)}=\frac{\beta_G\sum_{h_m}\sum_{h_n}(\beta_L\gamma)^{h_m+h_n}\gamma^{(h_m+h_n)g_t}p(h_m)p(h_n)\ \sum_{g_t}(\beta_G)^{g_t}p(g_t)}{\sum_{h_m}\sum_{h_n}(\beta_L)^{h_m+h_n}\gamma^{(h_m+h_n)g_t}p(h_m)p(h_n)\ \sum_{g_t}\beta_G^{g_t}p(g_t)}.$$

To calculate the power, we assume the allele frequency at locus L and G is 0.01 and 0.05, respectively. We first derive $\mathrm{P}(\mathrm{H}_S,\mathrm{H}_{NS}|\mathrm{AA}_\mathrm{R})$ then calculate the expectation of H_S and H_{NS} given both affteced siblings. The penetrance for one individual given genotype at loci L and G is $P(A|\,L_1=l,G_1=g)=\beta_0\beta_L^l\beta_G^g\gamma^{l\times g}$ where $L_1=h_m+h_n$ and $G_1=g_s+g_t$ denote the number of risk alleles at loci L and G, respectively. Define $vns_1\in\{0,1,2\}$ and $vns_2\in\{0,1,2\}$ as the number of non-shared allele on the first and second sibling, respectively. Let $vs\in\{0,1,2\}$ be the number of shared alleles between a sibpair. The probability of having both affected siblings given the number of shared and non-shared alleles is

$$\begin{aligned}P(AA|vns_1,vns_2,vs)&\\&=\sum_{G_1=0}^{2}\sum_{G_2=0}^{2}P(A|L_1=vns_1+vs,G_1)P(A|L_2=vns_2+vs,G_2)\\&\times\sum_{S=0}^{2}P(S)P(G_2|G_1,S)P(G_1)\,,\end{aligned}$$

then the expectation of $\mathrm{H}_S=vs$ and $\mathrm{H}_{NS}=vns_1+vns_2$ are

$$\mathrm{E}(\mathrm{H_{NS}}|\mathrm{AA})=\sum_{vns_1}\sum_{vns_2}\sum_{vs}(vns_1+vns_2)P(vns_1,vns_2,vs|AA)$$

$$\mathrm{E}(\mathrm{H_S}|\mathrm{AA})=\sum_{vns_1}\sum_{vns_2}\sum_{vs}(vs)P(vns_1,vns_2,vs|AA)$$

where

$$P(vns_1, vns_2, vs|AA) = \frac{P(vns_1, vns_2, vs, AA)}{P(AA)}$$

$$P(vns_1, vns_2, vs, AA) = P(AA|vns_1, vns_2, vs)P(vns_1, vns_2, vs)$$

$$= P(AA|vns_1, vns_2, vs)\sum_{S=0}^{2} P(vns_1, vns_2, vs|S)P(S)$$

To calculate the number of shared and non-shared chromosome regions, the derivation is similar but has to be conditional on the number of shared chromosome region S at locus L. Assuming no linkage between L and G, we integrate out the contribution from the remaining genome. The expectation of shared and non-shared chromosome can be calculated based on,

$$P(S|\ AA) = \frac{1}{\mathrm{P(AA)}}\sum_{vns_1}\sum_{vns_2}\sum_{vs} P(AA|vns_1, vns_2, vs, S)P(vns_1, vns_2, vs|S)P(S)$$

where

$$P(AA|vns_1, vns_2, vs, S)$$

$$= \sum_{G_1=0}^{2}\sum_{G_2=0}^{2} \mathrm{P}(A|L_1 = vns_1 + vs, G_1, S)P(A|L_2 = vns_2 + vs, G_2, S)$$

$$\times \sum_{S^*=0}^{2} P(S^*)P(G_2|G_1, S^*)P(G_1)$$

where S^* is the sharing status at locus G.

The sibling relative risk (SRR) is defined as

$$\mathrm{SRR} = \frac{P(AA)}{P(A)P(A)}$$

where

$P(A) = \sum_{L_1=0}^{2}\sum_{G_1}^{2} P(A|L_1, G_1)P(L_1, G_1) = \sum_{L_1=0}^{2}\sum_{G_1}^{2} P(A|L_1, G_1)P(L_1)P(G_1)$,

$P(AA) = \sum_{L_1=0}^{2} \sum_{G_1}^{2} P(AA|L_1, G_1)P(L_1, G_1)$, and

$$\Pr(\mathrm{AA}|L_1, G_1, S)$$

$$= P(A|L_1, G_1) \sum_{L_2=0}^{2} \sum_{G_2=0}^{2} P(A|L_2, G_2)P(L_2|L_1, S)\,(P(S^* = 0)P(G_2|G_1, S^* = 0)$$

$$+ P(S^* = 1)P(G_2|G_1, S = 1) + P(S^* = 2)P(G_2|G_1, S^* = 2))$$

where we assume no linkage between loci L and G, and S^* is the sharing status at locus G. Given the marginal relative risk of locus L, we find the marginal relative risk of locus G that satisfies the specified level of SRR assuming no gene-gene interaction ($\gamma = 1$) by solving β_L and β_G.

A.6. Population stratification in loci with linkage signal

In known loci, the linkage between the tested variants and the true risk variant could potentially cause the population stratification as a confounder using TRAFIC. To evaluate the extent of how linkage affects TRAFIC's robustness to population stratification, we perform simulations with different levels of linkage signal (LOD score) and show the genomic control λ under the null that the tested variant is not associated with the disease.

Define $\lambda_{IBD=2} = \lambda_{MZ}$ is the recurrence risk for a sibpair who share 2 IBD chromosomes, $\lambda_{IBD=1} = \lambda_O$ is the recurrence risk for a sibpair who share 1 IBD chromosome, $\lambda_{IBD=0} = 1$ is the recurrence risk for a sibpair who share 0 IBD chromosome. Then $\lambda_S = \frac{1}{4}(\lambda_{MZ} + 2\lambda_O + 1)$ is the recurrence risk for a sibpair. Let

z_0, z_1, z_2 be the proportion of sampled sibpairs who share 0,1,2 IBD chromosome, respectively. The expected values for z_0, z_1, z_2 is,

$$E(z_0) = 0.25\frac{1}{\lambda_S}, E(z_1) = 0.50\frac{\lambda_O}{\lambda_S}, E(z_2) = 0.25\frac{\lambda_{MZ}}{\lambda_S}$$

Assuming N sibpairs, the expected number of sibpairs who share 0, 1, and 2 IBD chromosomes are n_{IBD_0}, n_{IBD_1}, and n_{IBD_2}, respectively, where $n_{IBD_0} = N\times E(z_0)$, $n_{IBD_1} = N\times E(z_1)$, $n_{IBD_2} = N\times E(z_2)$. Thus, the expected LOD score is,

$$LOD = \log_{10}\frac{L(E(z_0),\ E(z_1), E(z_2))}{L(z_0 = \frac{1}{4}, z_1 = \frac{1}{2}, z_2 = \frac{1}{4})} = \log_{10}\frac{E(z_0)^{n_{IBD_0}}\times E(z_1)^{n_{IBD_1}}\times E(z_2)^{n_{IBD_2}}}{0.25^{n_{IBD_0}}\times 0.50^{n_{IBD_1}}\times 0.25^{n_{IBD_2}}}$$

We simulate 500 sibpairs from each population; one population has LOD score at the specified level and the other population with LOD = 0. Then, we apply TRAFIC and calculate the Genomic Control λ. Genomic Control λ grows with increasing LOD as shown in Table A-2. In addition, based on the expected values of z_0, z_1, z_2 and the allele frequencies of 0.01 and 0.05 in two populations, we calculate the analytical false positive rate at $\alpha = 0.05$.

LOD score	Genomic Control λ	False Positive Rate
3	1.06	0.057
5	1.13	0.063
10	1.30	0.081
57	5.73	0.352

Table A-2. The Genomic Control λ and false positive rate under different LOD score using TRAFIC. Calculations are based on a summed allele frequency of 0.01 in population 1 with the specified LOD and a summed allele frequency of 0.05 in population 2 with LOD = 0. The sample size of 1000 sibpairs are draw evenly from two populations.

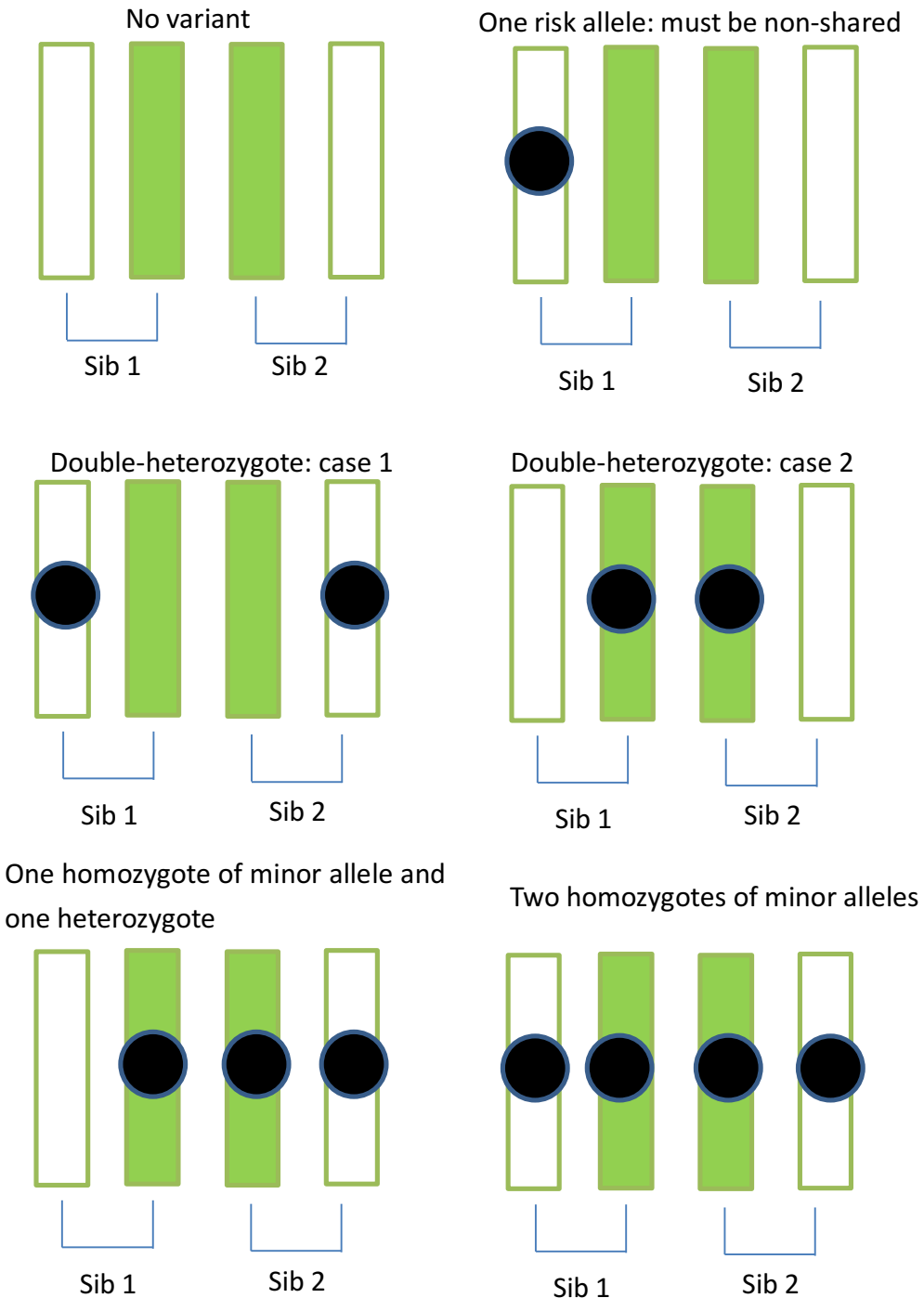


Figure A-1. Illustration of all possible sharing scenarios between a sibpair who shares one IBD chromosome region. When both siblings are heterozygotes, it is not possible to distinguish whether the minor allele is located on shared or non-shared chromosome regions. Block dot represents a minor allele, a colored rectangle is a shared IBD chromsome and a blank rectangle is a non-IBD chromosome region.

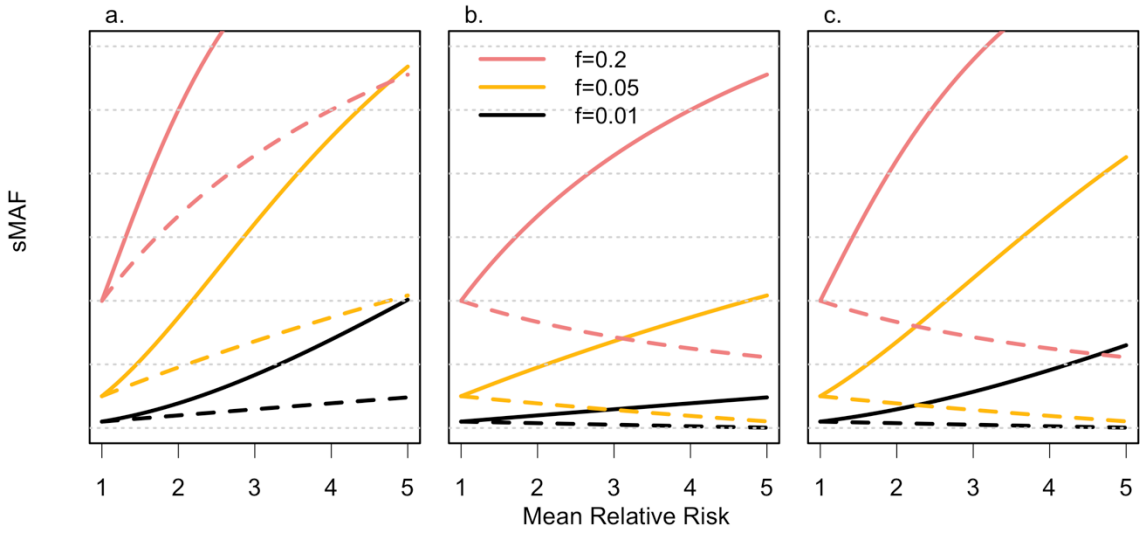


Figure A-2. Summed minor allele frequency (sMAF) of risk variants in cases (solid lines) and controls (broken lines) under different study designs at the disease prevalence of 0.20. We show sMAF as a function of mean relative risk of risk variants for (a) TRAFIC, (b) the conventional case-control design, and (c) the selected cases design for summed allele frequencies (f) of 0.01, 0.05 and 0.2 and fixed variance of relative risk $\sigma^2 = 0$.

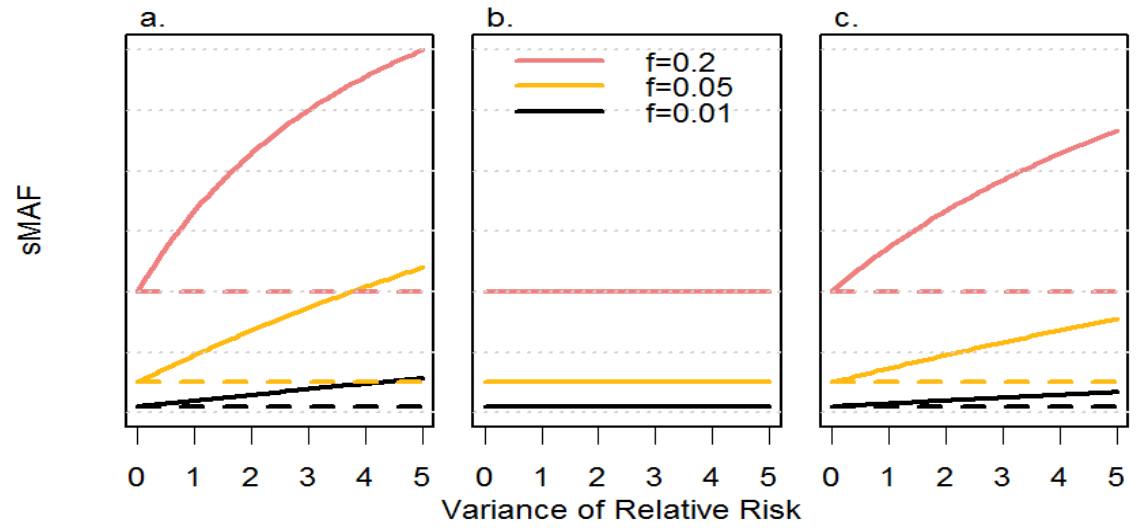


Figure A-3. Summed minor allele frequency (sMAF) of risk variants in cases (solid lines) and controls (broken lines) under different study designs at the disease prevalence of 0.01. We show sMAF as a function of variance of relative risk between risk variants for (a) TRAFIC, (b) the conventional case-control design, and (c) the selected cases design for summed allele frequencies (f) of 0.01, 0.05 and 0.2 and fixed mean relative risk $\mu = 1$.

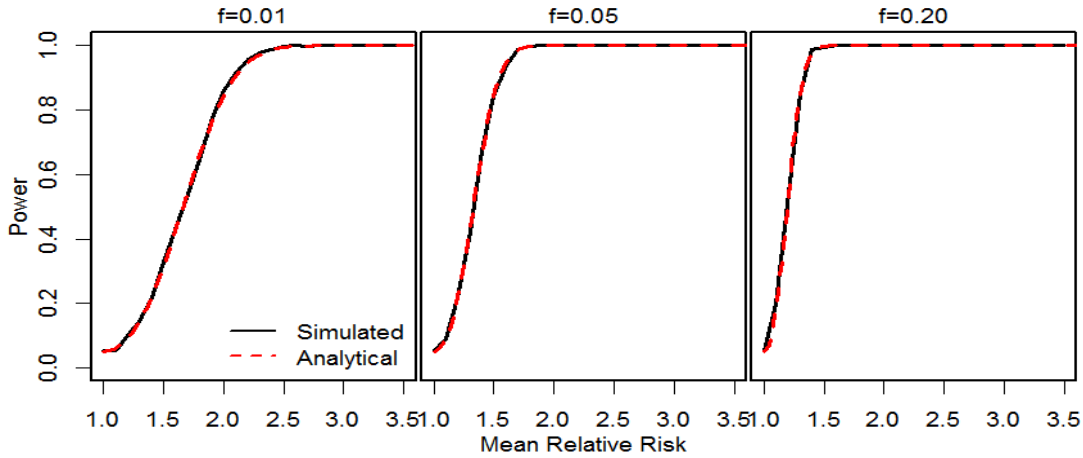


Figure A-4. The simulated and analytical power curve TRAFIC under different summed allele frequency f. The analytical and simulated power lines are in red and black, respectively. The power is evaluated at $\sigma^2 = 0$ while varying mean relative risk assuming the sharing status of alleles is known. Results are shown for 1000 sibpairs at a false positive rate of 0.05.

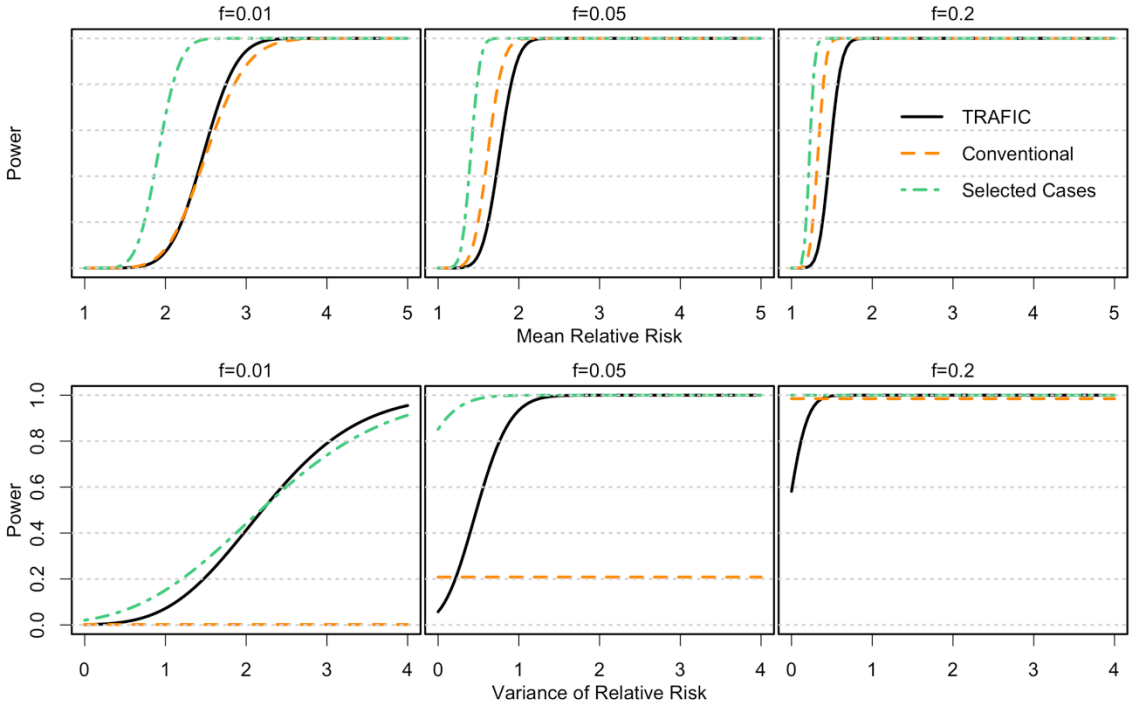


Figure A-5. The analytical power curve for TRAFIC, conventional case-control, and selected cases design for different summed allele frequencies (f) at the disease prevalence of 0.20. Row (a) displays the power as a function of mean relative risk evaluated at variance of relative risk $\sigma^2 = 0$. Row (b) shows the power as a function of variance of relative risk evaluated at mean relative risk $\mu = 1.5$. Results are shown for 2000 individuals (1000 sibpairs or 1000 cases and 1000 controls) at a significance level 2.5×10^{-6}.

Appendix B
For Chapter 3

B.1. Verification of Lyapunov condition in Lyapunov central limit theorem

First we describe Lyapunov central limit theorem. Let X_i be a sequence of independent random variables, each with finite expected value μ_i and variance σ_i^2

Define,

$$S_n^2 = \sum_{i=1}^{n} \sigma_i^2$$

If for some $\delta > 0$, the Lyapunov's condition holds,

$$\lim_{n\to\infty} \frac{1}{S_n^{2+\delta}} \sum_{i=1}^{n} E\left[|X_i - \mu_i|^{2+\delta}\right] = 0$$

Then,

$$\frac{1}{S_n} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{d} N(0,1)$$

To proof the Lyapunov condition holds, I consider families with two-generation (two founders and two offspring: one affected founder and both affected offspring). For all combinations of the number of shared IBD chromosomes between two siblings and the

number of carrier chromosomes (C) in founders, I calculated the expectation and variance for each combination. Assuming f = 0.01, since majority of informative families carrying only one carrier founder chromosome, I only consider the informative families which contains one carrier chromosome in the founders (IBD0C1, IBD1C1, and IBD2C1).

	IBD0C1	IBD1C1	IBD2C1	IBD0C2	IBD1C2	IBD2C2	IBD0C3	IBD1C3	IBD2C3
proportion	0.00975	0.0195	0.00975	0.00015	0.00029	0.00015	9.9e-07	1.98e-06	9.9e-07
Expectation	1.5	1.5	1.5	3	3	3	4.5	4.5	4.5
Variance	0.25	0.75	1.25	0.333	1	1.667	0.25	0.75	1.25
Delta=1	0.125	0.9375	1.75	0.333	1	3	0.125	0.9375	1.75

Non-informative	C0	C4
proportion	0.961	1e-08

For $\delta = 1$, the Lyapunov's condition is,

$$\frac{1}{S_n^{2+\delta}} \sum_{i=1}^{n} E\left[|X_i - \mu_i|^{2+\delta}\right]$$

$$= \frac{N_{fam} * (0.00975 * 0.125 + 0.0195 * 0.9375 + 0.00975 * 1.75)}{\left(N_{fam} * (0.00975 * 0.25 + 0.0195 * 0.75 + 0.00975 * 1.25)\right)^{\frac{3}{2}}}$$

$$= \frac{N_{fam} * (0.0365625)}{\left(N_{fam} * (0.02925)\right)^{\frac{3}{2}}} = \frac{0.0365625}{\sqrt{N_{fam}}(0.02925)^{\frac{3}{2}}} \to 0 \; as \; N_{fam} \to \infty$$

where $S_n^2 = \sum_{i=1}^{n} \sigma_i^2$. For different pedigree structures, I can use the similar approach to proof that Lyapunov condition holds.

B.2. Inheritance vector and missing founder imputation

B.2.1 Impute inheritance vector

Here, we demonstrate the imputation procedure with three affected offspring as in Figure B-1, however, the imputation works for any pedigree structure. As the IBD sharing status is known between siblings, the possible transmission pattern can be determined conditional on the IBD status. Without loss of generosity, the unobserved founder chromosomes are labelled as A, B, C, D; A, B are paternal chromosomes and C, D are maternal chromosomes. Assume we know the IBD sharing status between siblings as shown below. At the beginning, we assign A, C to Sib 1. Since Sib 1 and Sib 2 share 2 IBD chromosome region, Sib 2 must also have A, C. Given Sib 3 share one IBD chromosome region with Sib 1 and Sib 2, Sib 3 could have A, D or B, C. The choice of possible chromosomes for Sib 3 is independent of the final imputation result since in this step A,B,C,D are simply arbitrary labels and help determine the shared and non-shared chromosomes. In this example, we choose B, C.

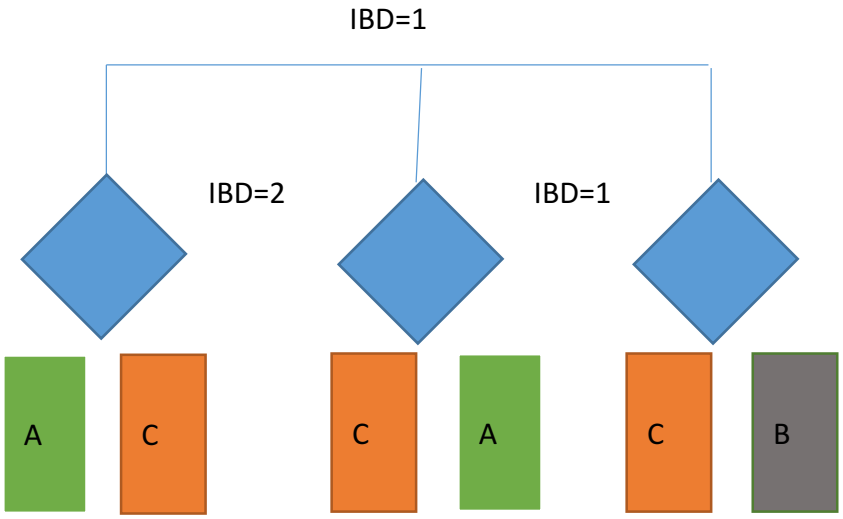


Figure B-1. The illustration of assigning a possible inheritance vector given sharing status between siblings.

Given inheritance vectors, phasing of variants are straightforward. When there are ambiguous phasing configurations, we impute the phasing of variants. Suppose we observe all three siblings are heterozygotes at this position, the phasing for this variant has two possible configurations; there is a single variant residing on chromosome C which is shared by three siblings or there are two variants which reside on shared chromosomes A and non-shared chromosome B. And we impute the phasing of the variants using EM algorithm, in which we estimate the minor allele frequencies on shared chromosome regions p_S and non-shared chromosome regions p_{NS} at a single SNP position. Across families, let N_S be the number of shared chromosome regions, x_S be the number of minor alleles located on IBD chromosome regions, N_{NS} be the number of non-IBD chromosome regions, and x_{NS} be the number of minor alleles located on non-IBD chromosome regions. Then x_S, x_{NS} are binomially distributed with the full likelihood function,

$$L(x_S, x_{NS}|N_S, N_{NS}, p_S, p_{NS}) = \binom{N_S}{x_S} p_S^{x_S}(1-p_S)^{N_S-x_S} \binom{N_{NS}}{x_{NS}} p_{NS}^{x_{NS}}(1-p_{NS})^{N_{NS}-x_{NS}}.$$

For unambiguous families, let k_S be the known total number of shared alleles and k_{NS} be the known total number of non-shared alleles. Suppose there are u ambiguous sibpairs. Without loss of generosity, for *i*th families with an ambiguous phase, u_{S_i} denotes number of shared allele and u_{NS_i} denotes non-shared allele. The likelihood function of observing the total number of shared allele x_S and non-shared allele x_{NS} can be rewritten as

$$\binom{N_S}{k_S+u_S} p_S^{k_S+u_S}(1-p_S)^{N_S-k_S-u_S} \binom{N_{NS}}{k_{NS}+2u_{NS}} p_{NS}^{k_{NS}+2u_{NS}}(1-p_{NS})^{N_{NS}-k_{NS}-2\mu_{NS}}$$

where $u_S = \sum_i u_{S_i}$ and $\mu_{NS} = \sum_i \mu_{NS_i}$ To set up the initial values for the algorithm, we estimate the allele frequency on IBD chromosome regions p_S by counting the total number of non-shared alleles x_S divided by the number of shared chromosomes N_S. Similarly, we estimate the allele frequency on non-IBD chromosome regions p_{NS}.

In the E-step, conditional on the current estimation of parameters, we calculate the expectation of shared and non-shared allele for *i*th ambiguous family weighted by each possible phase configuration. Both configurations contain one shared allele, thus

$$E\left(u_{S_i}|p_S, p_{NS}\right) = \frac{p_S(1-p_s)(1-p_{NS})}{p_S(1-p_s)(1-p_{NS}) + p_S(1-p_s)p_{NS}} + \frac{p_S(1-p_s)p_{NS}}{p_S(1-p_s)(1-p_{NS}) + p_S(1-p_s)p_{NS}}$$

Only the second configuration contain one non-shared allele, thus

$$E\left(u_{NS_i}|p_S, p_{NS}\right) = \frac{p_S(1-p_s)p_{NS}}{p_S(1-p_s)(1-p_{NS}) + p_S(1-p_s)p_{NS}}.$$

In the M-step, we update new p_S and p_{NS} as the solution of maximum likelihood estimators of the expected likelihood function

$$p_S^{new} = \frac{k_S + u_S}{N_S}, \quad p_{NS}^{new} = \frac{k_{NS} + 2u_{NS}}{N_{NS}}$$

where $u_S = \sum_i u_{S_i}$ and $\mu_{NS} = \sum_i \mu_{NS_i}$.The algorithm repeats between E-step and M-step until the estimation of parameters converges.

To impute the phase status for ambiguous families, let p_1 and p_2 be the probability of the first and the second phase configuration respectively.

$$p_1 = \frac{p_S^*(1-p_S^*)(1-p_{NS}^*)}{p_S^*(1-p_S^*)(1-p_{NS}^*) + p_S^*(1-p_S^*)p_{NS}^*},$$
$$p_2 = \frac{p_S^*(1-p_S^*)p_{NS}^*}{p_S^*(1-p_S^*)(1-p_{NS}^*) + p_S^*(1-p_S^*)p_{NS}^*}$$

where p_S^* and p_{NS}^* are the final estimates of EM. Then we can assign a phasing based on these probabilities.

B.2.2 Impute the carrier status of missing founder chromosomes

After assigning the inheritance vectors, we use multiple imputation to account for the uncertainty of estimating $p^{carrier}$ and impute founder genotypes to apply TRAP. We consider three imputation procedures as shown in Appendix B.3 and chose the procedure with the best power (Combine_Test). The imputation procedure has the following steps:

1. Estimate $\hat{p}^{carrier}$ using all samples and consider estimation uncertainty

   Naïve estimate from samples is $\hat{p}^{carrier} = \frac{\sum_{i=1}^{n_{fam}} g_i^{obs}}{\sum_{i=1}^{n_{fam}} n_i^{obs}}$

   g_i^{obs}: number of observed carrier parental chromosomes

   n_i^{obs}: number of observed parental chromosomes

   n_{fam}: number of families

2. Instead of using naïve estimator, we account for estimation uncertainty and draw

$\hat{p}^{carrier} \sim \text{Beta}(1 + \sum_{i=1}^{n_{fam}} g_i^{obs}, 1 + \sum_{i=1}^{n_{fam}} n_i^{obs} - \sum_{i=1}^{n_{fam}} g_i^{obs})$

3. Impute the number of carrier chromosomes in missing founders. For example, suppose for family *i*, there are $n_i^{missing}$ missing founder chromosomes, we impute the number of carrier chromosomes $x_i^{carrier}$ based on $x_i^{carrier} \sim Binomial\left(n_i^{missing}, \hat{p}^{carrier}\right)$ and calculate μ_i and σ_i^2 for each family
4. Repeat the procedures 2 and 3 for *M*=50 times
5. Combine M imputed data of each family then apply TRAP (Combine_Test)

$$\overline{\mu_i} = \sum_{l=1}^{M} \frac{\mu_i^{(l)}}{M}$$

$$v_i^2 = \frac{\sum_{l=1}^{M} \sigma_i^{2(l)}}{M} + \frac{M+1}{M} \frac{1}{M-1} \sum_{l=1}^{M} (\mu_i^{(l)} - \overline{\mu_i})^2$$

$$Test\ Statistic = \frac{\sum_i^{n_{fam}} (T_i - \overline{\mu_i})}{\sqrt{\sum_{i=1}^{n_{fam}} v_i^2}} \sim N(0,1)$$

B.3. Compare three combining rules for multiple imputation for TRAP

Suppose I use M rounds of multiple imputation, let i be the i^{th} family among n_{fam} families and T_i be the observed count of carrier chromosome among affected siblings. Given the l^{th} round of imputed data, I calculated the expectation $\mu_i^{(l)}$ and variance $\sigma_i^{2(l)}$ for each family. I compared the following three combing rules,

1. Combine M imputed data of each family then apply TRAP (Combine_Test)

$$\overline{\mu_i} = \sum_{l=1}^{M} \frac{\mu_i^{(l)}}{M}$$

$$v_i^2 = \frac{\sum_{l=1}^{M} {\sigma_i}^{2(l)}}{M} + \frac{M+1}{M}\frac{1}{M-1}\sum_{l=1}^{M}(\mu_i^{(l)} - \overline{\mu_i})^2$$

$$Test\ Statistic = \frac{\sum_i^{n_{fam}}(T_i - \overline{\mu_i})}{\sqrt{\sum_{i=1}^{n_{fam}} v_i^2}} \sim N(0,1)$$

2. Apply TRAP then combine across multiple imputation (Test_Combine)

$$D^{(l)} = \frac{\sum_{i=1}^{n_{fam}}(T_i - \mu_i^{(l)})}{\sqrt{\sum_{i=1}^{n_{fam}} {\sigma_i}^{2(l)}}}$$

$$\overline{D} = \frac{\sum_{l=1}^{M} D^{(l)}}{M}$$

$$V_D = 1 + \frac{M+1}{M}\frac{1}{M-1}\sum_{l=1}^{M}(D^{(l)} - \overline{D})^2$$

$$test\ statistic = \frac{\overline{D}}{\sqrt{V_D}} \sim N(0,1)$$

3. Apply TRAP then combine across multiple imputation with Chi-square statistic (Test_Combine_Chisq)

$$X^{(l)} = D^{2(l)} = \frac{(\sum_{i=1}^{n_{fam}}(T_i - \mu_i^{(l)}))^2}{\sum_{i=1}^{n_{fam}} {\sigma_i}^{2(l)}}$$

$$\overline{X} = \frac{\sum_{i=1}^{M} X^{(l)}}{M}$$

$$a = \frac{\sum_{i=1}^{M} \sqrt{X^{(l)}}}{M}$$

$$b = \frac{\sum_{l=1}^{M}(\sqrt{X^{(l)}} - a)^2}{M-1}$$

$$r_M = \frac{M+1}{M} b$$

$$v_M = (M-1)(1+r_M^{-2})$$

$$test\ statistic = \frac{\overline{X} - \frac{(M+1)}{M-1} r_M}{1+r_M} \sim F_{1,v_M}$$

I used simulations to evaluate the performance. I simulated 1000 families with two parents and three children, and consider randomly mask both parents in 100%, 50%, and 20% of families. To quantify the power loss, I compared imputed results to the result without missing founder (noimpute). Type I error was set at 0.05.

B.3.1 Power comparison for 100% missing parents

Considering all parents were missing, combine_test strategy performed slightly better than test_combine. Test_combine_chisq has the least power.

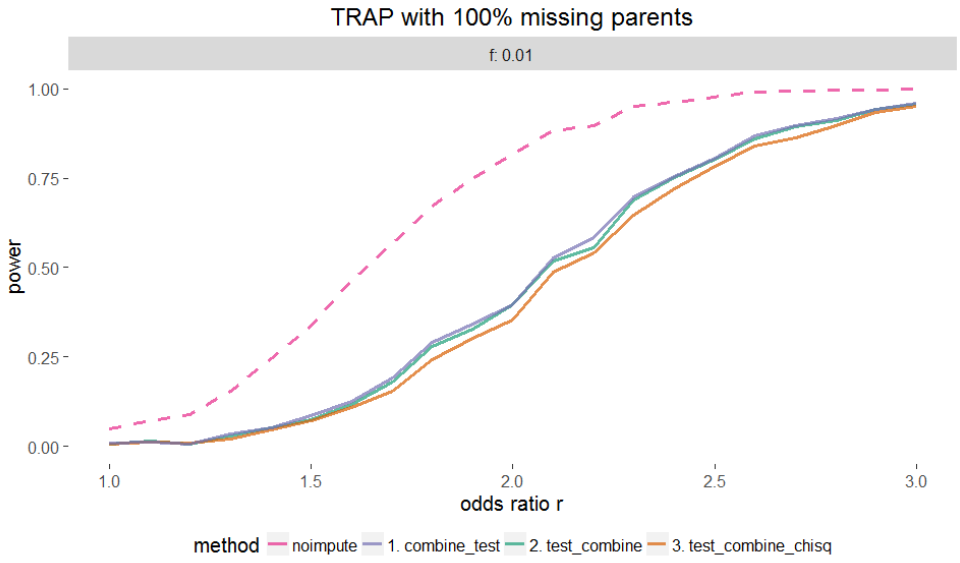


Figure B-2. Power comparison across three imputation procedures with 100% missing parents and evaluated at $\alpha = 0.05$.

B.3.2 Power comparison for 50% missing parents

When there were 50% of families with missing founders, the difference across three combining rules reduced.

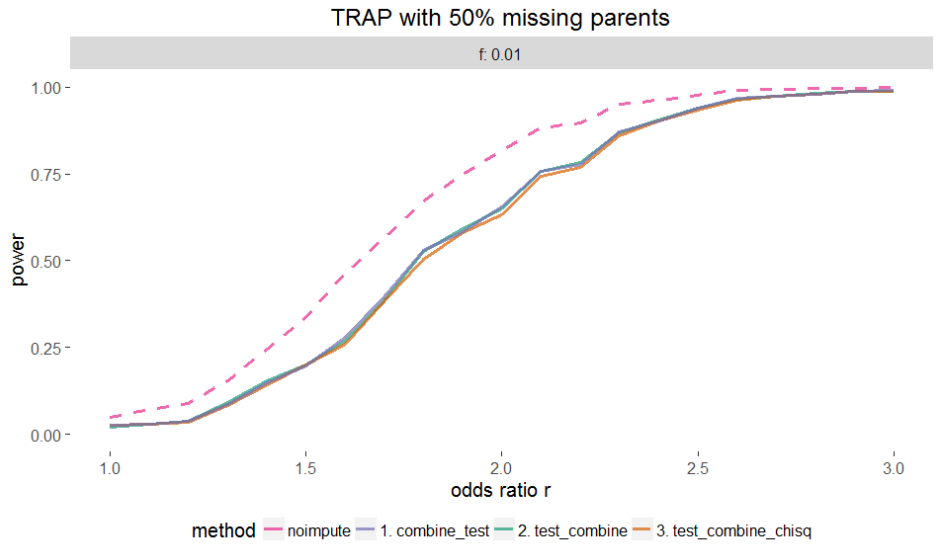


Figure B-3. Power comparison across three imputation procedures with 50% missing parents and evaluated at $\alpha = 0.05$.

B.3.3 Power comparison for 20% missing parents

When there were only 20% of families with missing parents, there was no significant difference across three combining rules.

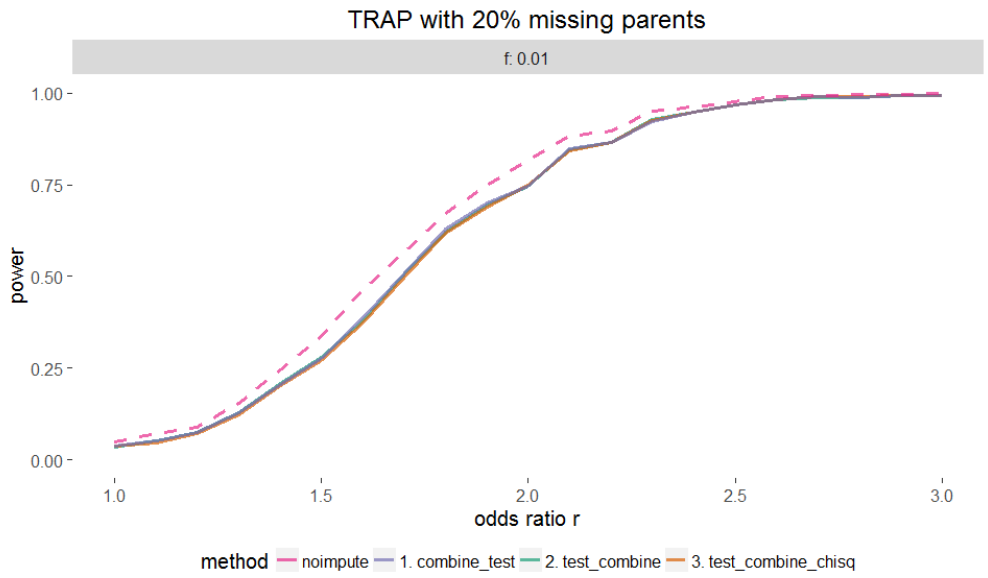


Figure B-4. Power comparison across three imputation procedures with 50% missing parents and evaluated at $\alpha = 0.05$.

All three combining rules were valid and conservative under the null with very similar performance in the considered scenarios. When all families had missing parents, the first combining rule (combine first then test) had a slightly advantage in power over the other two rules. When only a minor portion of families with missing parents, all three combining rules performed well.

B.4. Additional Figures

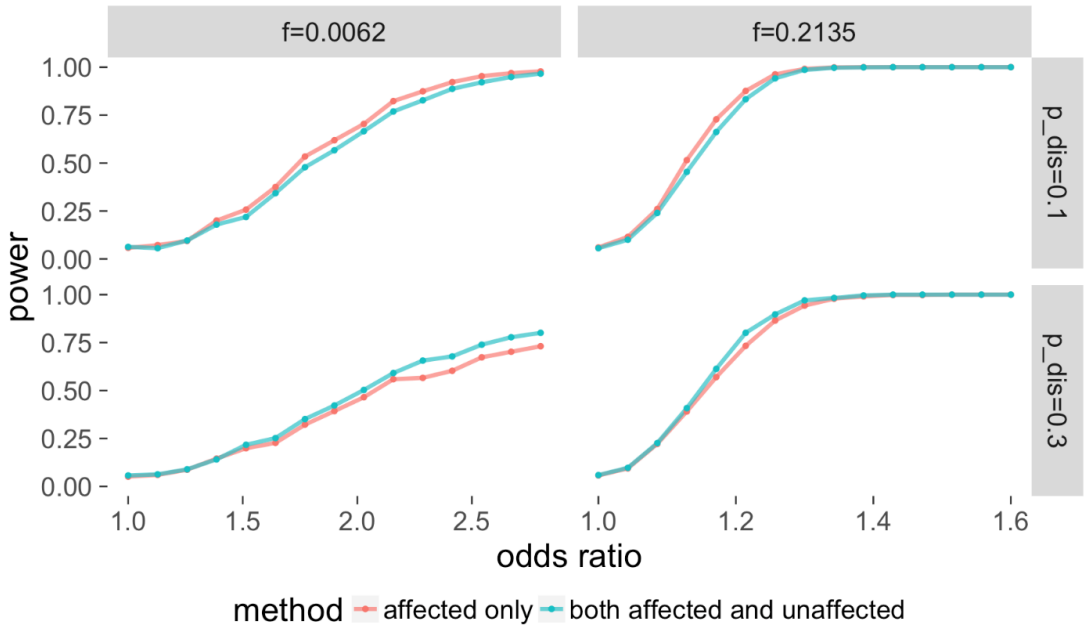


Figure B-5. Power comparison between full samples and affected-only designs for TRAP evaluated at $\alpha = 0.05$. Power curve is shown as a function of effect size r (odds ratio) of risk variants in a gene with risk chromosome frequency f and disease prevalence p_dis. The pedigree structure is two-founder-two-children (one affected founder and two affected children)

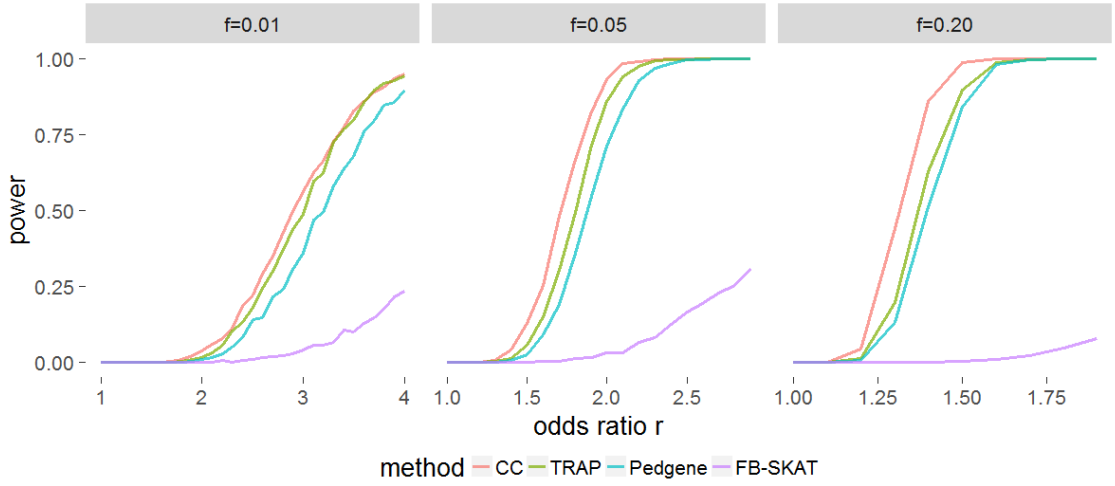


Figure B-6. Power comparison across TRAP, Pedgene, FB-SKAT, and case-control design (CC) evaluated at $\alpha = 2.5\times10^{-6}$ and prevalence 10%. Power curve is shown as a function of effect size r (odds ratio) of risk variants in a gene with carrier chromosome frequency f = 0.01, 0.05, 0.20.

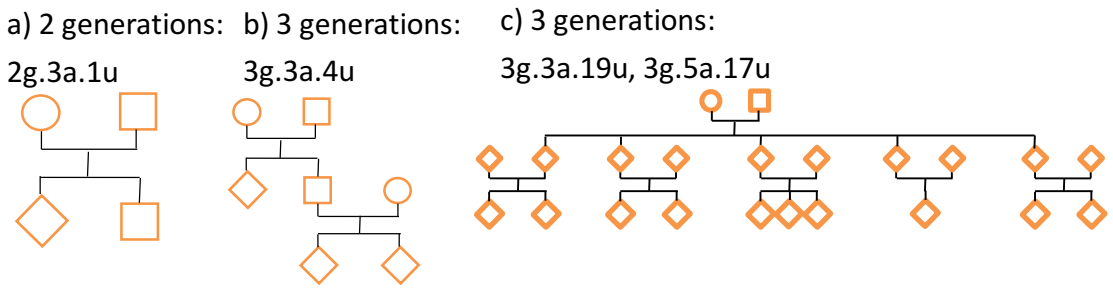


Figure B-7. Illustration of number of members in each generation for different family structures. "g" denotes number of generations, "a" denotes number of affected individuals that can occur in any generation, and "u" denotes the number of unaffected individuals.

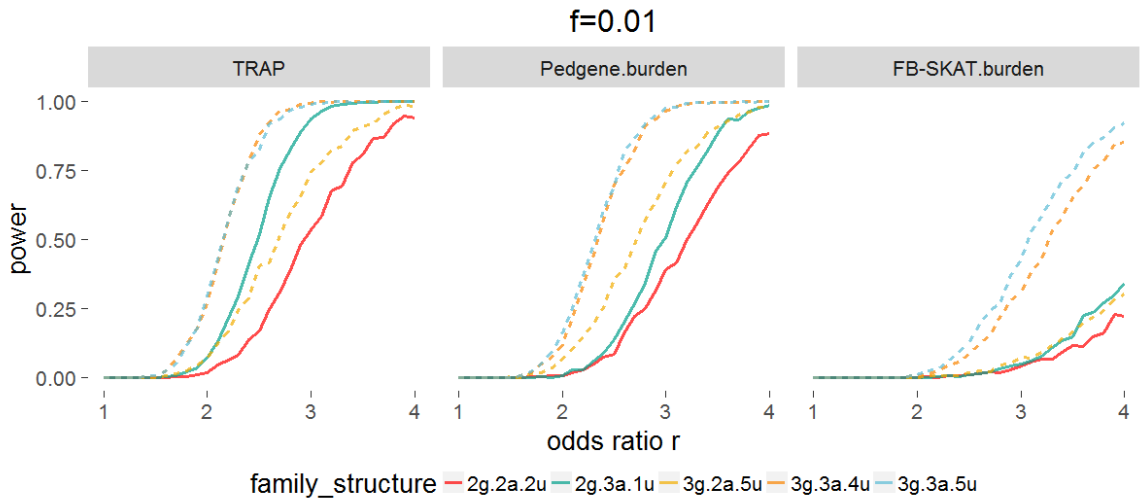


Figure B-8. Power comparison across pedigree structures for TRAP, Pedgene and FB-SKAT evaluated at $\alpha = 2.5\times10^{-6}$ *and f = 0.01. The power curve is shown as a function of odds ratio of risk variants.*

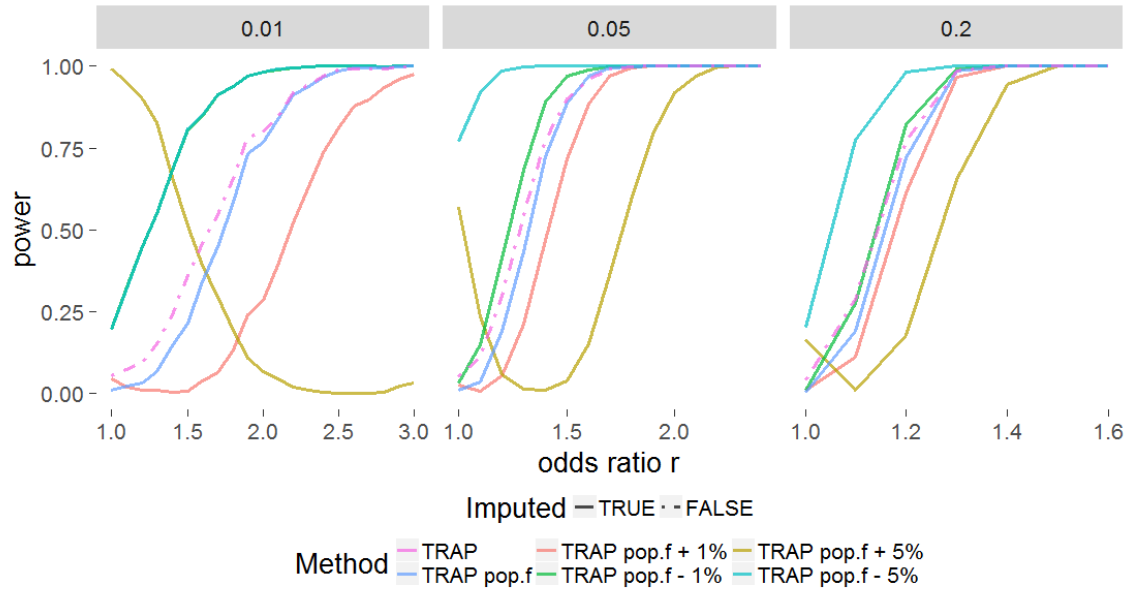


Figure B-9. Power curve for TRAP using imputaion for missing founders and using external allele frequency, evlaued at $\alpha = 0.05$. The imputation algorithm assuming carrier chromosome frequency from an existing database (e.g. 1000 Genome Peoject) "pop.f" with the absolute bias in allele ferquency.

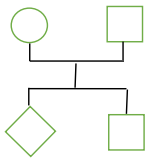
	f=0.01
Complete data -- No imputation	1.4×10^{-4}
Imputed samples with 20% missing founders	0.6×10^{-4}
Imputed samples with 50% missing founders	0.2×10^{-4}
Imputed samples with 80% missing founders	0.01×10^{-4}
Imputed samples with 100% missing founders	$< 0.01\times10^{-4}$

Table B-1. Type I error rate evaluation for TRAP with and without imputed samples under different population carrier frequency given $\alpha = 2.5\times10^{-4}$.

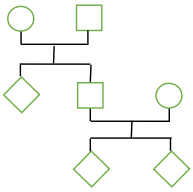
Appendix C
For Chapter 4

C.1. Pedigree structures illustration

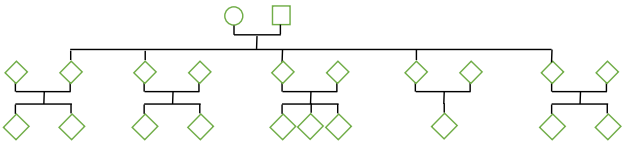
a) 2 generations:
2g.4n, 2g.3a.1u



b) 3 generations:
3g.7n, 3g.3a.4u

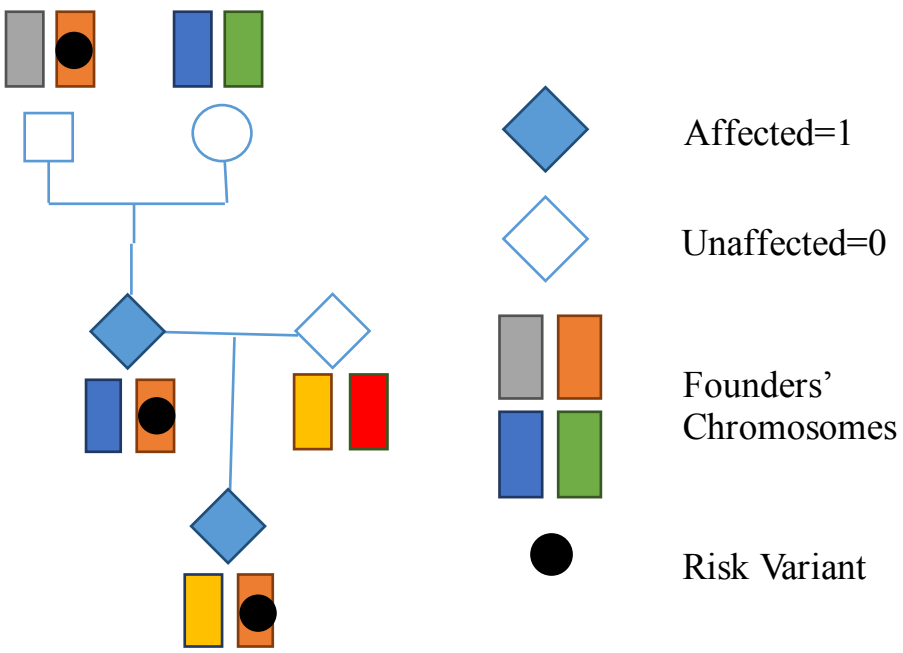


c) 3 generations:
3g.22n, 3g.3a.19u, 3g.5a.17u



C.2. Investigate how TRAP, Pedgene, and FB-SKAT use within-family information

Given pedigree structure as below,



To compare how TRAP, Pedgene, FB-SKAT use within-family information, I consider TRAP in two scenarios: first look at each trio separately, equivalent to using FB-SKAT. Second, I use the regular TRAP to analyze this pedigree. I also compare to Pedgene by assuming a simple burden test. I quantify the overall information by their resulting Z statistics.

When separately evaluating each trio, the enumerated summary statistics for each trio based on TRAP (FB-SKAT) is (0,1,1,0), thus $\mu = 0.5$ and $\sigma^2 = 0.25$. Thus, the overall information is (1-0.5+1-0.5)/sqrt(0.25+0.25)=1.414

When collectively looking at the pedigree, the enumerated summary statistics for each trio based on TRAP is (0,2,1,0,1,0), thus $\mu = 0.67$ and $\sigma^2 = 0.83$. The overall information is (2-0.67)/sqrt(0.83)=1.459

I evaluated Pedgene using the following equation,

$$\text{Pedgene\_burden} = T_i - E(T_i) = \sum_{j}^{n_i} (Y_{ij} - \overline{Y_i})(X_{ij} - 2f)$$

$$\text{Var}(\text{ Pedgene\_burden}) = \sum_{j}^{n_i} \sum_{j'}^{n_i} 2f(1-f)(Y_{ij} - \overline{Y_i})(Y_{ij'} - \overline{Y_i})\text{Cov}(X_{ij}, X_{ij'}))$$

where $\text{Cov}(X_{ij}, X_{ij'}) = \begin{cases} 0 & IBD = 0 \\ 0.5 & IBD = 1 \\ 1 & IBD == 2 \end{cases}$

I used the estimated frequency $\hat{f}$ based on founders. In this case, Pedgene_burden=0.8, Var(Pedgene_burden)=0.58, and the overall information is 1.05. Based on the overall information, TRAP is the most efficient method.

In conclusion, under ascertained scenarios, between-family information is not useful since every family has a similar mean phenotype and mean genotype. Thus, within-family information is the most informative. Under non-ascertained scenarios, between-information is useful. But, for large pedigree, within family information is more informative with a smaller number of families. Thus, within-family information out-weight the between-family information, and TRAP can exploit the sharing among all family members to outperform Pedgene.

Bibliography

Abecasis, Gonçalo R., Lon R. Cardon, and William O. Cookson. 2000. “A General Test of Association for Quantitative Traits in Nuclear Families.” *American Journal of Human Genetics* 66 (1): 279–92. doi:10.1086/302698.

Abecasis, Gonçalo R., Stacey S. Cherny, William O. Cookson, and Lon R. Cardon. 2002. “Merlin--Rapid Analysis of Dense Genetic Maps Using Sparse Gene Flow Trees.” *Nature Genetics* 30 (1): 97–101. doi:10.1038/ng786.

Altmüller, Janine, Lyle J. Palmer, Guido Fischer, Hagen Scherb, and Matthias Wjst. 2001. “Genomewide Scans of Complex Human Diseases: True Linkage Is Hard to Find.” *American Journal of Human Genetics* 69 (5): 936–50. doi:10.1086/324069.

Altshuler, David, Mark J. Daly, and Eric S. Lander. 2008. “Genetic Mapping in Human Disease.” *Science (New York, N.Y.)* 322 (5903): 881–88. doi:10.1126/science.1156409.

Ashley, Euan A. 2016. “Towards Precision Medicine.” *Nature Reviews Genetics* 17 (9): 507–22. doi:10.1038/nrg.2016.86.

Bodea, Corneliu A., Benjamin M. Neale, Stephan Ripke, International IBD Genetics Consortium, Mark J. Daly, Bernie Devlin, and Kathryn Roeder. 2016. “A Method to Exploit the Structure of Genetic Ancestry Space to Enhance Case-Control Studies.” *American Journal of Human Genetics* 98 (5): 857–68. doi:10.1016/j.ajhg.2016.02.025.

Burton, Paul R., David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, et al. 2007. “Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls.” *Nature* 447 (7145): 661–78. doi:10.1038/nature05911.

Chen, Han, James B. Meigs, and Josée Dupuis. 2013. “Sequence Kernel Association Test for Quantitative Traits in Family Samples.” *Genetic Epidemiology* 37 (2): 196–204. doi:10.1002/gepi.21703.

Chen, Han, Chaolong Wang, Matthew P. Conomos, Adrienne M. Stilp, Zilin Li, Tamar Sofer, Adam A. Szpiro, et al. 2016. “Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models.” *American Journal of Human Genetics* 98 (4): 653–66. doi:10.1016/j.ajhg.2016.02.012.

Chen, Rong, Lisong Shi, Jörg Hakenberg, Brian Naughton, Pamela Sklar, Jianguo Zhang, Hanlin Zhou, et al. 2016. “Analysis of 589,306 Genomes Identifies Individuals Resilient to Severe Mendelian Childhood Diseases.” *Nature Biotechnology* 34 (5): 531–38. doi:10.1038/nbt.3514.

Chen, Wei-Min, and Gonçalo R. Abecasis. 2007. “Family-Based Association Tests for Genomewide Association Scans.” *American Journal of Human Genetics* 81 (5): 913–26.

Cirulli, Elizabeth T., and David B. Goldstein. 2010. “Uncovering the Roles of Rare Variants in Common Disease through Whole-Genome Sequencing.” *Nature Reviews Genetics* 11 (6): 415–25. doi:10.1038/nrg2779.

Conomos, Matthew P., Alexander P. Reiner, Bruce S. Weir, and Timothy A. Thornton. 2016. “Model-Free Estimation of Recent Genetic Relatedness.” *American Journal of Human Genetics* 98 (1): 127–48. doi:10.1016/j.ajhg.2015.11.022.

Cooper, Gregory M., and Jay Shendure. 2011. “Needles in Stacks of Needles: Finding Disease-Causal Variants

in a Wealth of Genomic Data." *Nature Reviews Genetics* 12 (9): 628–40. doi:10.1038/nrg3046.

De, Gourab, Wai-Ki Yip, Iuliana Ionita-Laza, and Nan Laird. 2013. "Rare Variant Analysis for Family-Based Design." *PloS One* 8 (1): e48495.

Devlin, Bernie, and Kathryn Roeder. 1999. "Genomic Control for Association Studies." *Biometrics* 55 (4): 997–1004. doi:10.1111/j.0006-341X.1999.00997.x.

Edwards, Stacey L., Jonathan Beesley, Juliet D. French, and Alison M. Dunning. 2013. "Beyond GWASs: Illuminating the Dark Road from Association to Function." *American Journal of Human Genetics* 93 (5): 779–97. doi:10.1016/j.ajhg.2013.10.012.

Eichler, Evan E., Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. 2010. "Missing Heritability and Strategies for Finding the Underlying Causes of Complex Disease." *Nature Reviews. Genetics* 11 (6): 446–50. doi:10.1038/nrg2809.

Epstein, Michael P., Richard Duncan, Erin B. Ware, Min A. Jhun, Lawrence F. Bielak, Wei Zhao, Jennifer A. Smith, Patricia A. Peyser, Sharon L. R. Kardia, and Glen A. Satten. 2015. "A Statistical Approach for Rare-Variant Association Testing in Affected Sibships." *American Journal of Human Genetics* 96 (4): 543–54. doi:10.1016/j.ajhg.2015.01.020.

Eu-ahsunthornwattana, Jakris, E. Nancy Miller, Michaela Fakiola, Selma M. B. Jeronimo, Jenefer M. Blackwell, Heather J. Cordell, and Wellcome Trust Case Control Consortium 2. 2014. "Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data." *PLoS Genet* 10 (7): e1004445. doi:10.1371/journal.pgen.1004445.

Fang, Shurong, Qiuying Sha, and Shuanglin Zhang. 2012. "Two Adaptive Weighting Methods to Test for Rare Variant Associations in Family-Based Designs." *Genetic Epidemiology* 36 (5): 499–507. doi:10.1002/gepi.21646.

Feng, Shuang, Giorgio Pistis, He Zhang, Matthew Zawistowski, Antonella Mulas, Magdalena Zoledziewska, Oddgeir L. Holmen, et al. 2015. "Methods for Association Analysis and Meta-Analysis of Rare Variants in Families." *Genetic Epidemiology* 39 (4): 227–38. doi:10.1002/gepi.21892.

Fingerlin, Tasha E., Michael Boehnke, and Gonçalo R. Abecasis. 2004. "Increasing the Power and Efficiency of Disease-Marker Case-Control Association Studies through Use of Allele-Sharing Information." *The American Journal of Human Genetics* 74 (3): 432–43. doi:10.1086/381652.

Fritsche, Lars G., Wilmar Igl, Jessica N. Cooke Bailey, Felix Grassmann, Sebanti Sengupta, Jennifer L. Bragg-Gresham, Kathryn P. Burdon, et al. 2016. "A Large Genome-Wide Association Study of Age-Related Macular Degeneration Highlights Contributions of Rare and Common Variants." *Nature Genetics* 48 (2): 134–43. doi:10.1038/ng.3448.

Fuchsberger, Christian, Jason Flannick, Tanya M. Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J. Gaulton, Clement Ma, et al. 2016. "The Genetic Architecture of Type 2 Diabetes." *Nature* 536 (7614): 41–47. doi:10.1038/nature18642.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." *Nature Reviews. Genetics* 17 (6): 333–51. doi:10.1038/nrg.2016.49.

Guan, Weihua, Michael Boehnke, Anna Pluzhnikov, Nancy J. Cox, and Laura J. Scott. 2012. "Identifying Plausible Genetic Models Based on Association and Linkage Results: Application to Type 2 Diabetes." *Genetic Epidemiology* 36 (8): 820–28. doi:10.1002/gepi.21668.

Guo, Wei, and Yin Yao Shugart. 2012. "Detecting Rare Variants for Quantitative Traits Using Nuclear Families." *Human Heredity* 73 (3): 148–58. doi:10.1159/000338439.

Hakonarson, Hakon, Struan F. A. Grant, Jonathan P. Bradfield, Luc Marchand, Cecilia E. Kim, Joseph T. Glessner, Rosemarie Grabs, et al. 2007. "A Genome-Wide Association Study Identifies KIAA0350 as a Type 1 Diabetes Gene." *Nature* 448 (7153): 591–94. doi:10.1038/nature06010.

Hampe, Jochen, Andre Franke, Philip Rosenstiel, Andreas Till, Markus Teuber, Klaus Huse, Mario Albrecht, et al. 2007. "A Genome-Wide Association Scan of Nonsynonymous SNPs Identifies a Susceptibility

Variant for Crohn Disease in ATG16L1." *Nature Genetics* 39 (2): 207–11. doi:10.1038/ng1954.

Helbig, Ingo, Susan E. Hodge, and Ruth Ottman. 2013. "Familial Cosegregation of Rare Genetic Variants with Disease in Complex Disorders." *European Journal of Human Genetics : EJHG* 21 (4): 444–50. doi:10.1038/ejhg.2012.194.

Howson, Joanna M. M., N. M. Walker, D. Clayton, J. A. Todd, and Type 1 Diabetes Genetics Consortium. 2009. "Confirmation of HLA Class II Independent Type 1 Diabetes Associations in the Major Histocompatibility Complex Including HLA-B and HLA-A." *Diabetes, Obesity & Metabolism* 11 Suppl 1 (February): 31–45. doi:10.1111/j.1463-1326.2008.01001.x.

Hunt, Kelly J., Donna M. Lehman, Rector Arya, Sharon Fowler, Robin J. Leach, Harald H. H. Göring, Laura Almasy, et al. 2005. "Genome-Wide Linkage Analyses of Type 2 Diabetes in Mexican Americans: The San Antonio Family Diabetes/Gallbladder Study." *Diabetes* 54 (9): 2655–62.

Hunter, David J. 2005. "Gene-Environment Interactions in Human Diseases." *Nature Reviews. Genetics* 6 (4): 287–98. doi:10.1038/nrg1578.

Ionita-Laza, Iuliana, Seunggeun Lee, Vladimir Makarov, Joseph D Buxbaum, and Xihong Lin. 2013. "Family-Based Association Tests for Sequence Data, and Comparisons with Population-Based Association Tests." *European Journal of Human Genetics* 21 (10): 1158–62. doi:10.1038/ejhg.2012.308.

Ionita-Laza, Iuliana, and Ruth Ottman. 2011. "Study Designs for Identification of Rare Disease Variants in Complex Diseases: The Utility of Family-Based Designs." *Genetics* 189 (3): 1061–68. doi:10.1534/genetics.111.131813.

Jiang, Yunxuan, Karen N. Conneely, and Michael P. Epstein. 2014. "Flexible and Robust Methods for Rare-Variant Testing of Quantitative Traits in Trios and Nuclear Families." *Genetic Epidemiology* 38 (6): 542–51. doi:10.1002/gepi.21839.

Kang, Hyun Min, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-Yee Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. 2010. "Variance Component Model to Account for Sample Structure in Genome-Wide Association Studies." *Nature Genetics* 42 (4): 348–54. doi:10.1038/ng.548.

Keith, Jonathan M., Allan McRae, David Duffy, Kerrie Mengersen, and Peter M. Visscher. 2008. "Calculation of IBD Probabilities with Dense SNP or Sequence Data." *Genetic Epidemiology* 32 (6): 513–19. doi:10.1002/gepi.20324.

Laird, Nan M., Steve Horvath, and Xin Xu. 2000. "Implementing a Unified Approach to Family-Based Tests of Association." *Genetic Epidemiology* 19 Suppl 1: S36-42. doi:10.1002/1098-2272(2000)19:1+<::AID-GEPI6>3.0.CO;2-M.

Laird, Nan M., and Christoph Lange. 2006. "Family-Based Designs in the Age of Large-Scale Gene-Association Studies." *Nature Reviews. Genetics* 7 (5): 385–94. doi:10.1038/nrg1839.

Lander, Eric S., and Nicholas J. Schork. 1994. "Genetic Dissection of Complex Traits." *Science (New York, N.Y.)* 265 (5181): 2037–48.

Lee, Seunggeung, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. 2014. "Rare-Variant Association Analysis: Study Designs and Statistical Tests." *American Journal of Human Genetics* 95 (1): 5–23. doi:10.1016/j.ajhg.2014.06.009.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616): 285–91. doi:10.1038/nature19057.

Li, Bingshan, and Suzanne M. Leal. 2008. "Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data." *American Journal of Human Genetics* 83 (3): 311–21. doi:10.1016/j.ajhg.2008.06.024.

Li, Xin, and Jing Li. 2011. "Haplotype Reconstruction in Large Pedigrees with Untyped Individuals through IBD Inference." *Journal of Computational Biology* 18 (11): 1411–21. doi:10.1089/cmb.2011.0167.

Lin, Keng-Han, and Sebastian Zöllner. 2015. "Robust and Powerful Affected Sibpair Test for Rare Variant Association." *Genetic Epidemiology* 39 (5): 325–333.

Liu, Qianying, Dan L. Nicolae, and Lin S. Chen. 2013. “Marbled Inflation from Population Structure in Gene-Based Association Studies with Rare Variants.” *Genetic Epidemiology* 37 (3): 286–92. doi:10.1002/gepi.21714.

MacCluer, Jean W., John L. VandeBerg, Bruce Read, and Oliver A. Ryder. 1986. “Pedigree Analysis by Computer Simulation.” *Zoo Biology* 5 (2): 147–60. doi:10.1002/zoo.1430050209.

Mahmood, Syed S., Daniel Levy, Ramachandran S. Vasan, and Thomas J. Wang. 2014. “The Framingham Heart Study and the Epidemiology of Cardiovascular Diseases: A Historical Perspective.” *Lancet* 383 (9921): 999–1008. doi:10.1016/S0140-6736(13)61752-3.

Mancuso, Nicholas, Nadin Rohland, Kristin A. Rand, Arti Tandon, Alexander Allen, Dominique Quinque, Swapan Mallick, et al. 2016. “The Contribution of Rare Variation to Prostate Cancer Heritability.” *Nature Genetics* 48 (1): 30–35. doi:10.1038/ng.3446.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. “Finding the Missing Heritability of Complex Diseases.” *Nature* 461 (7265): 747–53. doi:10.1038/nature08494.

Mathieson, Iain, and Gil McVean. 2012. “Differential Confounding of Rare and Common Variants in Spatially Structured Populations.” *Nature Genetics* 44 (3): 243–46. doi:10.1038/ng.1074.

McCarthy, Mark I., Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. 2008. “Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges.” *Nature Reviews Genetics* 9 (5): 356–69. doi:10.1038/nrg2344.

McQueen, Matthew B., B. Devlin, Stephen V. Faraone, Vishwajit L. Nimgaonkar, Pamela Sklar, Jordan W. Smoller, Rami Abou Jamra, et al. 2005. “Combined Analysis from Eleven Linkage Studies of Bipolar Disorder Provides Strong Evidence of Susceptibility Loci on Chromosomes 6q and 8q.” *American Journal of Human Genetics* 77 (4): 582–95. doi:10.1086/491603.

Nelson, Matthew R., Toby Johnson, Liling Warren, Arlene R. Hughes, Stephanie L. Chissoe, Chun-Fang Xu, and Dawn M. Waterworth. 2016. “The Genetics of Drug Efficacy: Opportunities and Challenges.” *Nature Reviews Genetics* 17 (4): 197–206. doi:10.1038/nrg.2016.12.

Nelson, Matthew R., Daniel Wegmann, Margaret G. Ehm, Darren Kessner, Pamela St Jean, Claudio Verzilli, Judong Shen, et al. 2012. “An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People.” *Science* 337 (6090): 100–104. doi:10.1126/science.1217876.

Ng, Sarah B., Kati J. Buckingham, Choli Lee, Abigail W. Bigham, Holly K. Tabor, Karin M. Dent, Chad D. Huff, et al. 2010. “Exome Sequencing Identifies the Cause of a Mendelian Disorder.” *Nature Genetics* 42 (1): 30–35. doi:10.1038/ng.499.

Ng, Sarah B., Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, et al. 2009. “Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes.” *Nature* 461 (7261): 272–76. doi:10.1038/nature08250.

O’Connell, Jared, Deepti Gurdasani, Olivier Delaneau, Nicola Pirastu, Sheila Ulivi, Massimiliano Cocca, Michela Traglia, et al. 2014. “A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness.” *PLOS Genet* 10 (4): e1004234. doi:10.1371/journal.pgen.1004234.

Ott, Jurg, Yoichiro Kamatani, and Mark Lathrop. 2011. “Family-Based Designs for Genome-Wide Association Studies.” *Nature Reviews. Genetics* 12 (7): 465–74. doi:10.1038/nrg2989.

Ott, Jurg, Jing Wang, and Suzanne M. Leal. 2015. “Genetic Linkage Analysis in the Age of Whole-Genome Sequencing.” *Nature Reviews. Genetics* 16 (5): 275–84. doi:10.1038/nrg3908.

Peng, Bo, Biao Li, Younghun Han, and Christopher I. Amos. 2010. “Power Analysis for Case-Control Association Studies of Samples with Known Family Histories.” *Human Genetics* 127 (6): 699–704.

Pericak-Vance, M. A. 2001. “Analysis of Genetic Linkage Data for Mendelian Traits.” *Current Protocols in Human Genetics* Chapter 1 (May): Unit 1.4. doi:10.1002/0471142905.hg0104s09.

Price, Alkes L., Gregory V. Kryukov, Paul I. W. de Bakker, Shaun M. Purcell, Jeff Staples, Lee-Jen Wei, and Shamil R. Sunyaev. 2010. “Pooled Association Tests for Rare Variants in Exon-Resequencing

Studies." *The American Journal of Human Genetics* 86 (6): 832–38. doi:10.1016/j.ajhg.2010.04.005.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9. doi:10.1038/ng1847.

Rakovski, Cyril S., Xin Xu, Ross Lazarus, Deborah Blacker, and Nan M. Laird. 2007. "A New Multimarker Test for Family-Based Association Studies." *Genetic Epidemiology* 31 (1): 9–17. doi:10.1002/gepi.20186.

Rao, Dabeeru C., Michael A. Province, Mark F. Leppert, A. l. Oberman, Gerardo Heiss, R. Curt Ellison, Donna K. Arnett, et al. 2003. "A Genome-Wide Affected Sibpair Linkage Analysis of Hypertension: The HyperGEN Network." *American Journal of Hypertension* 16 (2): 148–50. doi:10.1016/S0895-7061(02)03247-8.

Risch, Neil. 1990. "Linkage Strategies for Genetically Complex Traits. I. Multilocus Models." *American Journal of Human Genetics* 46 (2): 222–28.

Risch, Neil, and K. Merikangas. 1996. "The Future of Genetic Studies of Complex Human Diseases." *Science (New York, N.Y.)* 273 (5281): 1516–17.

Roach, Jared C., Gustavo Glusman, Robert Hubley, Stephen Z. Montsaroff, Alisha K. Holloway, Denise E. Mauldin, Deepak Srivastava, et al. 2011. "Chromosomal Haplotypes by Genetic Phasing of Human Families." *American Journal of Human Genetics* 89 (3): 382–97. doi:10.1016/j.ajhg.2011.07.023.

Schaffner, Stephen F., Catherine Foo, Stacey Gabriel, David Reich, Mark J. Daly, and David Altshuler. 2005. "Calibrating a Coalescent Simulation of Human Genome Sequence Variation." *Genome Research* 15 (11): 1576–83. doi:10.1101/gr.3709305.

Schaid, Daniel J., Shannon K. McDonnell, Jason P. Sinnwell, and Stephen N. Thibodeau. 2013. "Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods with Pedigree or Population Structured Data." *Genetic Epidemiology* 37 (5): 409–18. doi:10.1002/gepi.21727.

Schifano, Elizabeth D., Michael P. Epstein, Lawrence F. Bielak, Min A. Jhun, Sharon L. R. Kardia, Patricia A. Peyser, and Xihong Lin. 2012. "SNP Set Association Analysis for Familial Data." *Genetic Epidemiology* 36 (8): 797–810. doi:10.1002/gepi.21676.

Sham, Pak C., Shaun Purcell, Stacey S. Cherny, and Gonçalo R. Abecasis. 2002. "Powerful Regression-Based Quantitative-Trait Linkage Analysis of General Pedigrees." *American Journal of Human Genetics* 71 (2): 238–53.

Shugart, Yin Yao, Yun Zhu, Wei Guo, and Momiao Xiong. 2012. "Weighted Pedigree-Based Statistics for Testing the Association of Rare Variants." *BMC Genomics* 13 (November): 667. doi:10.1186/1471-2164-13-667.

Sidore, Carlo, Fabio Busonero, Andrea Maschio, Eleonora Porcu, Silvia Naitza, Magdalena Zoledziewska, Antonella Mulas, et al. 2015. "Genome Sequencing Elucidates Sardinian Genetic Architecture and Augments Association Analyses for Lipid and Blood Inflammatory Markers." *Nature Genetics* 47 (11): 1272–81. doi:10.1038/ng.3368.

Sladek, Robert, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, et al. 2007. "A Genome-Wide Association Study Identifies Novel Risk Loci for Type 2 Diabetes." *Nature* 445 (7130): 881–85. doi:10.1038/nature05616.

Spielman, Richard S., Ralph E. McGinnis, and Warren J. Ewens. 1993. "Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM)." *American Journal of Human Genetics* 52 (3): 506–16.

Sul, Jae Hoon, Brian E. Cade, Michael H. Cho, Dandi Qiao, Edwin K. Silverman, Susan Redline, and Shamil Sunyaev. 2016. "Increasing Generality and Power of Rare-Variant Tests by Utilizing Extended Pedigrees." *American Journal of Human Genetics* 99 (4): 846–59. doi:10.1016/j.ajhg.2016.08.015.

Tenesa, Albert, and Chris S. Haley. 2013. "The Heritability of Human Disease: Estimation, Uses and Abuses." *Nature Reviews Genetics* 14 (2): 139–49. doi:10.1038/nrg3377.

The 1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." *Nature*

526 (7571): 68–74. doi:10.1038/nature15393.

Uricchio, Lawrence H., Noah A. Zaitlen, Chun Jimmie Ye, John S. Witte, and Ryan D. Hernandez. 2016. “Selection and Explosive Growth Alter Genetic Architecture and Hamper the Detection of Causal Rare Variants.” *Genome Research* 26 (7): 863–73. doi:10.1101/gr.202440.115.

Wang, Chaolong, Xiaowei Zhan, Jennifer Bragg-Gresham, Hyun Min Kang, Dwight Stambolian, Emily Y. Chew, Kari E. Branham, et al. 2014. “Ancestry Estimation and Control of Population Stratification for Sequence-Based Association Studies.” *Nature Genetics* 46 (4): 409–15. doi:10.1038/ng.2924.

Wang, Xuefeng, Seunggeun Lee, Xiaofeng Zhu, Susan Redline, and Xihong Lin. 2013. “GEE-Based SNP Set Association Test for Continuous and Discrete Traits in Family-Based Association Studies.” *Genetic Epidemiology* 37 (8): 778–86. doi:10.1002/gepi.21763.

Weeks, Daniel E., and Kenneth Lange. 1988. “The Affected-Pedigree-Member Method of Linkage Analysis.” *American Journal of Human Genetics* 42 (2): 315–26.

Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, et al. 2014. “The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations.” *Nucleic Acids Research* 42 (Database issue): D1001-1006. doi:10.1093/nar/gkt1229.

Wheeler, Heather E., Michael L. Maitland, M. Eileen Dolan, Nancy J. Cox, and Mark J. Ratain. 2013. “Cancer Pharmacogenomics: Strategies and Challenges.” *Nature Reviews. Genetics* 14 (1): 23–34. doi:10.1038/nrg3352.

Whittemore, A. S., and J. Halpern. 1994. “A Class of Tests for Linkage Using Affected Pedigree Members.” *Biometrics* 50 (1): 118–27.

Wu, Michael C., Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. 2011. “Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test.” *American Journal of Human Genetics* 89 (1): 82–93. doi:10.1016/j.ajhg.2011.05.029.

Yang, Jian, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna A. E. Vinkhuyzen, Sang Hong Lee, Matthew R. Robinson, et al. 2015. “Genetic Variance Estimation with Imputed Variants Finds Negligible Missing Heritability for Human Height and Body Mass Index.” *Nature Genetics* 47 (10): 1114–20. doi:10.1038/ng.3390.

Zawistowski, Matthew, Shyam Gopalakrishnan, Jun Ding, Yun Li, Sara Grimm, and Sebastian Zöllner. 2010. “Extending Rare-Variant Testing Strategies: Analysis of Noncoding Sequence and Imputed Genotypes.” *The American Journal of Human Genetics* 87 (5): 604–17. doi:10.1016/j.ajhg.2010.10.012.

Zawistowski, Matthew, Mark Reppell, Daniel Wegmann, Pamela L. St Jean, Margaret G. Ehm, Matthew R. Nelson, John Novembre, and Sebastian Zöllner. 2014. “Analysis of Rare Variant Population Structure in Europeans Explains Differential Stratification of Gene-Based Tests.” *European Journal of Human Genetics* 22 (9): 1137–44. doi:10.1038/ejhg.2013.297.

Zhang, Qunyuan, Lihua Wang, Dan Koboldt, Ingrid B. Boreki, and Michael A. Province. 2014. “Adjusting Family Relatedness in Data-Driven Burden Test of Rare Variants.” *Genetic Epidemiology* 38 (8): 722–27. doi:10.1002/gepi.21848.

Zhang, Yiwei, Xiaotong Shen, and Wei Pan. 2013. “Adjusting for Population Stratification in a Fine Scale with Principal Components and Sequencing Data.” *Genetic Epidemiology* 37 (8). doi:10.1002/gepi.21764.

Zhou, Xiang, and Matthew Stephens. 2012. “Genome-Wide Efficient Mixed-Model Analysis for Association Studies.” *Nature Genetics* 44 (7): 821–24. doi:10.1038/ng.2310.

Zöllner, Sebastian. 2012. “Sampling Strategies for Rare Variant Tests in Case-Control Studies.” *European Journal of Human Genetics : EJHG* 20 (10): 1085–91.