Distribution Agreement

Signature:

______________________________ ______________

Yunxuan Jiang Date

Statistical Methods for Rare-Variant Sequencing Studies in Pedigrees

By

Yunxuan Jiang
Doctor of Philosophy

Biostatistics

Karen Conneely, Ph.D.
Advisor

Michael P. Epstein, Ph.D.
Advisor

Stephanie Sherman, Ph.D.
Committee Member

Yijuan Hu, Ph.D.
Committee Member

Zhaohui “Steve” Qin, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for Rare-Variant Sequencing Studies in Pedigrees

By

Yunxuan Jiang

M.S.P.H, Emory University, 2011
B.S., Beijing Forestry University, 2009

Advisors: Karen Conneely, PhD and Michael P. Epstein, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

Abstract

Next-generation sequencing studies have the potential to increase understanding of genetic architecture of complex diseases in more depth than ever before, but require the development of robust and powerful statistical methods to identify trait-influencing variation. During the past few years, interests have shifted from identifying common susceptibility variation in the population to rare susceptibility variation. However, the infrequent observation of rare variants (<5% in the population) poses difficulties in developing powerful statistical methods. Although methods have been proposed to analyze rare susceptibility variation in population-based or case-control designs, few of these methods can be applied to family-based study designs. Family-based designs have several advantages including higher power due to increased genetic load, robustness to population stratification, and the ability to identify de-novo mutations by sequencing trios. In our first project, we developed a flexible and robust method for rare variant analysis of quantitative traits in nuclear families and trios. Our method uses a kernel-machine framework to analyze rare variants in aggregate, and has the advantages of analytical calculation of p-values and robustness to population stratification. The method also employs a screening step to improve power. This method, as with other existing methods, mainly focuses on trios and nuclear families while ignoring the information provided by other types of relatives. As more studies tend to re-sequence subjects from previous linkage analysis studies, which normally involve more than two generations, statistical methods to analyze sequencing studies of large pedigrees are needed. In our second project, we develop a method for family-based rare-variant analysis of quantitative outcomes that can accommodate any family structure and size. Our first and second projects are designed to perform family-based tests that consider association between a gene and a single phenotype. However, there has been increasing interest in identifying pleiotropic genes through joint testing of multiple phenotypes; such approaches are both biologically meaningful and statistically more powerful than univariate testing of individual phenotypes. Therefore, in our third project, we develop a cross-phenotype association test for case-parent trio studies. Based on the kernel distance covariance framework, our test can incorporate multiple traits (both binary and continuous in nature) and is more powerful compared to analogous univariate tests of individual phenotypes.

Statistical Methods for Rare-Variant Sequencing Studies in Pedigrees

By

Yunxuan Jiang

M.S.P.H, Emory University, 2011
B.S., Beijing Forestry University, 2009

Advisors: Karen Conneely, PhD and Michael P. Epstein, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2017

Acknowledgement

I would like to express my heartfelt gratitude to my advisors Dr. Karen Conneely and Dr. Michael Epstein for their guidance, encouragement, and inspiration during the past seven years. They have been very generous with their time and help since my first day at Emory. They also set up role models for me and showed me the type of researchers I would like to be. I also would like to thank my committee members, Drs. Stephanie Sherman, Yijuan Hu and Steve Qin for their constructive suggestions, which significantly improved the quality of my work. Last, I would like to thank my parents for their unconditional love and support.

Contents

List of Figures

Chapter 1. Introduction

1.1 Background

The human genome is the code of our life; it affects traits like hair color and height, and, more importantly, it affects one's risk of disease. Because of this, genomics is playing a more and more important role in public health research. According to the Centers for Disease Control and Prevention (CDC), genetic factors are associated with nine of the ten leading causes of death in the United States, including heart disease, cancer, diabetes, and Alzheimer disease (https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm).

Over the past decade, researchers used genome wide association studies (GWAS) to identify genetic susceptibility variants that were common (population frequency > 5%) under the common-disease common-variant hypothesis (CDCV: Complex disease is attributable to a moderate number of common variants, each of modest effect on disease risk). This kind of research identified over 5000 SNP-trait associations for over 600 traits as of 2013 (Welter, MacArthur et al. 2014). However, most of these associated SNPs have very small effect sizes (odds ratio between 1.1-1.5), and the proportion of heritability (proportion of phenotypic variance in a population attributable to additive genetic factors) explained by these SNPs is at best modest for most traits. For example, current findings can only explain 6% of the heritability of type 2 diabetes (Manolio, Collins et al. 2009, Mathieson and McVean 2012). Given common susceptibility variants fail to explain a large proportion of heritability for most complex diseases, interest has shifted toward identifying disease-associated rare variants with population allele frequencies < 5%. Recent development of cost-effective sequencing technologies, especially next-generation sequencing methods, has made direct sequencing and analysis of rare variants feasible.

Association analysis of rare susceptibility variants is hampered by the fact that power to detect a risk variant is often limited since power generally decreases with a decrease in variant frequency when sample size and effect size are held constant. Because of this, standard analytic methods used for GWAS (e.g. Cochran-Armitage trend test) lose substantial power in rare variant sequencing studies. In my dissertation, I will focus on solving this issue by developing powerful and robust statistical methods for analysis of rare trait-influencing variation. Specifically, my methods will focus on methods for analyzing rare variants collected in family-based studies of complex traits.

Family-based designs have several advantages over population-based designs in that they enable the use of statistics that, by design, are robust to confounding due to population stratification. Population stratification is a well-known problem that can cause inflated false positive rates and decreased power to detect real association. Stratification arises from a systematic difference in allele frequencies between subjects sampled from different populations whose disease prevalence and allele frequencies are significantly different from each other. This is because each population has a unique social and genetic background; and social or cultural events, such as the mating process, will greatly influence the genetic architecture of a population (Cardon and Palmer 2003). Marchini et al. (Marchini, Cardon et al. 2004) showed that with the large sample size required for GWAS, even a small fraction of admixture between different populations will lead to a greatly inflated type 1 error rate, which increases rapidly as the sample size grows large or the population structure becomes more extreme. In our earlier work (Jiang, Epstein et al. 2013), we showed that for rare variant sequencing studies in case-control designs, type 1 error rate could increase up to 0.9 at alpha=0.05 when population structure is extreme. As a result, it is very important to adjust for population stratification or develop a method, which by design, is robust to population stratification. Using family-based designs, it is possible to create robust tests that are insensitive to population stratification.

Family designs also can solve genetic problems that are hard to answer in population-based studies. For example, sequencing the parents of affected subjects can identify *de novo* mutations and also allow the study of rare homozygous genotypes, which are difficult to find in population-based designs (Do, Kathiresan et al. 2012). Families are also attractive to study because they often provide increased genetic load for a disease or trait: while carriers of a rare risk allele will be hard to sample in the general population, they are more likely to be found in families of probands (Zollner 2012). Finally, family studies allow the study of the segregation pattern of complex disease (Ott, Kamatani et al. 2011). Because of these appealing features and the fact that there are many familial samples from past linkage studies that are available, family-based resequencing studies are gaining in popularity. Several recent studies have identified disease-associated rare variants through family-based designs, including rare variants associated with multiple sclerosis (Ramagopalan, Dyment et al. 2011), simplex autism (Krumm, O'Roak et al. 2013), dilated cardiomyopathy (Norton, Li et al. 2011), and Alzheimer's disease (Cruchaga, Haller et al. 2012).

In the remainder of this chapter, I will review literature on analytic methods for rare-variant analysis of complex traits, discuss outstanding issues in the area, and then provide some general detail about how the proposed work in this proposal for family-based analysis fills an important gap in the literature. I then provide a concluding paragraph describing an outline of subsequent chapters of this proposal.

1.2 Literature review

1.2.1 Existing methods for rare variant analysis

As the power of statistical methods decreases as allele frequency of a tested variant decreases, there has been general consensus that, rather than analyze rare variants individually, a test that groups or aggregates rare variants in a gene or region for analysis is likely optimal. These aggregate approaches can be broadly categorized as either burden tests that collapse grouped rare variants into a single aggregate variable that is then regressed on phenotype (Kwee, Liu et al. 2008, Madsen and Browning 2009, Morris and Zeggini 2010, Zawistowski, Gopalakrishnan et al. 2010), or kernel machine regression tests that relate phenotype to rare variants in a region as a function of a variance component (Wu, Lee et al. 2011). This aggregation strategy also extends to joint testing of multiple phenotypes together to identify rare variants in genes that are pleiotropic in nature. In the presence of pleiotropy, joint testing of these phenotypes is not only biological meaningful but also statistically more powerful than univariate analysis of each separate phenotype accounting for multiple testing. In subsequent paragraphs, we describe both univariate and cross-phenotype tests for rare-variant sequencing studies in detail.

1.2.1.1 Statistical methods for testing individual phenotype

Burden Tests The central idea of burden tests is to collapse all the rare variants in a region together into a composite variable (e.g. number of copies of rare variants a subject possesses in a gene), and then test the association between the composite variable and the disease. The disadvantages of the collapsing method are that it combines the functional and nonfunctional variants together and is sensitive to misclassification. Li and Leal polished this method by developing a “Combined Multivariate and Collapsing” (CMC) method (Li and Leal 2008). It collapses the rare variants in a region to several composite variables according to whether the variant is functional or not, and then does a multiple marker test. Their simulation results show that the CMC method has high power to identify disease associated rare variants and is sensitive to misclassification Madsen and Browning (Madsen and Browning 2009) improved this method by giving more weight to rare variants in a collapsed group that are rarer in the population (with the idea that the rarer the variant, the more likely it is causal since it is selected against in the population due to its deleterious nature). Price et al. (Price, Kryukov et al. 2010) developed a pooling method by assigning different thresholds to collapse variants for different genes. The idea of their method is that the effect size of a variant is not necessarily inversely proportional to the allele frequency in all situations. Instead of arbitrarily pooling alleles based on a specific minor-allele threshold, their pooling method analyzes the data multiple times across various frequency thresholds and then uses permutations to accommodate multiple testing.

The burden methods work well when all causal variants affect the outcome in the same direction (i.e. all causal variants increase risk or all causal variants decrease risk). However, the association of the composite variable will be attenuated when causal variants in a gene act in different directions on phenotype. In addition, the majority of variants typically have null or negligible effects on the outcome; when the causal variants are collapsed with these null variants, the power to identify the causal variant will also be attenuated. To deal with a potential sparsity of signal in a gene-based test, as well as the possibility that variants might act in different directions on phenotype, a new class of variance-component methods for rare-variant testing emerged, which we describe in the next section.

Kernel Machine Regression: Wu et al. (Wu, Lee et al. 2011) proposed a kernel-machine regression procedure for rare-variant association test, which they named the sequential kernel association test (SKAT) method. The SKAT method is a variance-component score test that is robust to direction of effects of the causal variants in the region and retains power when the signal among rare variants in a gene is sparse. They model the outcome as follows:

$$g(\mu_i) = \alpha_0 + \alpha' X_i + \beta' G_i \qquad (1)$$

where μ_i is the expected value of outcome y_i, $g(\mu_i) = \mu_i$ for continuous traits and $g(\mu_i) = \log it(p(y_i = 1))$ for binary traits, α_0 is the intercept term, X_i is the vector of covariates for individual i, and α' is a vector of the corresponding coefficients. G_i is a vector of genotypes for *p* variants within the region, where components in G_i represent allele counts for each variant and take the value of 0, 1, or 2 under an additive model (one can also use a dominant or recessive model, if desired). To test whether variation within the region significantly associates with the outcome or not, traditional omnibus methods require *p* degrees of freedom to test $\beta_1 = \beta_2 = \beta_3 = \beta_4 ...... = \beta_p = 0$. This type of test has very low power, especially for rare variant tests. To overcome this issue, Wu et al.'s method assumes that for any *j*, β_j follows an arbitrary distribution with mean zero and variance $w_j \tau$, where w_j is the pre-specified weight (with such weights, like the method of Madsen and Browning, being based on minor-allele frequency). It can be seen that testing whether rare-variants are associated with phenotype in SKAT is equivalent to testing whether the variance parameter τ is equal to zero. SKAT adopts a score test of the variance component $H_0 : \tau = 0$, which is shown to be the locally most powerful test by Lin (Lin 1997). The test statistic takes the form:

$$Q = (y - \hat{\mu})' K (y - \hat{\mu}) \qquad (2),$$

$K = GWG$ is the kernel function, which measures the genetic similarity between subjects, and $W = diag(w_{1,} w_{2,} ...., w_p)$ is the weight matrix. w_j can take different values based on pre-specified knowledge about the variant in the region. If all the w_j=1, then this is a linear kernel. Several other kernels are available for different situations, and a correctly specified kernel can help to increase power (Wu, Lee et al. 2011). Under the null hypothesis, the test statistic follows a mixture of chi-square distributions and one can analytically calculate p-values using Davies' method (Davies 1980), with no need for permutation. In addition to being powerful, the SKAT test is also computationally efficient, as it only needs to fit under the null model:

$$g(m_i) = a_0 + a' X_i \qquad (3)$$

In summary, SKAT has the advantages of power over a burden test when a region contains rare variants that act in different directions on phenotype. Additionally, SKAT is computationally efficient as it analytically derives p-values with no need for permutation. The method can also incorporate prior knowledge of the test region to increase power.

1.2.1.2 Statistical methods for joint testing of multiple phenotypes

Pleiotropy refers to the situation when one gene affects multiple phenotypes. For example, mutations in the Phenylalanine Hydroxylase Gene (PAH gene) are associated with mental retardation, eczema, and pigment defects(Paul 2000). Therefore, joint testing of phenotypes is biologically meaningful when analyzing pleiotropic genes. In addition, joint tests can increase effective sample size and thus gain power over univariate tests that consider each phenotype separately. In this section, we review existing methods for testing pleiotropic genes in rare-variant sequencing study when one consider gene-based testing of rare variants.

Kernel distance covariance test of independence Broadaway et al. (Broadaway, Cutler et al. 2016) proposed a cross-phenotype test based on the kernel distance covariance framework (Gretton, Fukumizu et al. 2007). The kernel independence test is a widely used approach in machine learning area to test the dependency of two random variables in the kernel space. After forming a kernel for each random variable, the method calculates the canonical correlation of two kernels. One important feature of this method is that the expectation of the cross-covariance of two kernels equals zero if and only if the underlying two random variables are independent (Bach and Jordan 2002). Despite the popularity of the test, it is hard to find an empirical estimate of the dependency criteria. To tackle this issue, Gretton et al. (Gretton, Fukumizu et al. 2007) proposed the use of Hilbert-Schmidt norm as a measure of the dependency. The idea is that if the correlation matrix equals zero then the sum of the square singular values of the matrix will also not deviate from zero.

Broadaway et al. (Broadaway, Cutler et al. 2016) leveraged KDC's framework to form a cross-phenotype test. Their method consists of two steps: they first construct a similarity matrix P for phenotypes and a similarity matrix K for genetic variants; they then form the test statistic as

$$Q \propto trace(KHPH), \tag{4}$$

where $\mathbf{H} = (\mathbf{I} - 1_N 1_N^T / N)$ is the centering matrix. Similar to KMR, choice of P and K depends on prior knowledge about the gene and the phenotypes. As shown in (4), the test statistic Q has a very simple form, which makes the calculation very straightforward. Asymptotically, Q follows a mixture of chi-square distribution thus p-values can be derived efficiently using Davies' method (Davies 1980). The method can also easily be adjusted for covariates by regressing each phenotype separately on the covariates and using the residuals to form the similarity matrix P. These characteristics make KDC ideal for testing pleiotropic genes in rare-variant sequencing study.

Besides KDC, another method that can be used for testing pleiotropic genes is multivariate kernel machine regression (MV-KMR(Maity, Sullivan et al. 2012). Similar to univariate KMR, MV-KMR can be reduced to a multivariate linear mixed model. Hua and Ghosh established the link between MV-KMR and KDC (Hua and Ghosh 2015). They showed that the test statistics of MV-KMR could also be written as:

$$Q \mu (Y - \bar{Y})K(Y - \bar{Y}) = tr[K(I - \frac{11'}{n})YY'(I - \frac{11'}{n})] = tr[KHYY'H] \tag{5}$$

Comparing (4) and (5), it can be shown that if the same K is used for two tests and linear kernel is used to form P in (4), then the two test statistics will have the same form. Theoretically, tests using KDC will achieve a least as much power as MV-KMR. If the correct phenotype similarity matrix is chosen, KDC will have greater than KMR.

1.2.2 Existing methods for family-based studies

Several methods have been proposed for univariate rare-variant association testing in families. Schaid et al. (Schaid, McDonnell et al. 2013) developed a method for complex traits that accounts for relatedness among study subjects. Their method took a retrospective view of the sample, which assumes that the outcome is fixed while the genotype is random, and is particularly appealing for the analysis of datasets that are collected under non-random ascertainment (such as those collected for linkage studies). Chen et al. (Chen, Meigs et al. 2013) developed a rare-variant test for quantitative traits in families by extending kernel-machine methods (Kwee, Liu et al. 2008, Wu, Lee et al. 2011) to pedigree analysis by inserting a random familial effect due to shared polygenes within the modeling framework; a similar idea was employed by Schifano et al. (Schifano, Epstein et al. 2012) and Oualkacha et al (Oualkacha, Dastani et al. 2013). Jiang et al. (Jiang and McPeek 2014) adopted a similar strategy and extended the SKAT-O (Lee, Emond et al. 2012) method to family studies of quantitative traits.

1.2.3 Population stratification and QTDT

Although the above methods for family-based studies account for correlation within families, they do not consider potential bias caused by population stratification. Population stratification can lead to substantially inflated false positive rates in sequencing studies of rare variants (Epstein, Duncan et al. 2012, Jiang, Epstein et al. 2013, Liu, Nicolae et al. 2013), and standard GWAS approaches to correct for such stratification (such as principal components or EMMAX (Kang, Sul et al. 2010)) may not be effective when applied to rare variants (Mathieson and McVean 2012). Therefore, a rare-variant association test that maintains validity in the presence of such stratification is needed. Ionita-Laza et al. (Ionita-Laza, Lee et al. 2013) proposed such a method based on the family-based association test (FBAT) framework. Although this method is robust to population stratification, it ignores between-family information that could perhaps be exploited to boost power. Fang et al. (Fang, Sha et al. 2012, Fang, Zhang et al. 2013) used between-family information for this purpose in an adaptive rare-variant association test for quantitative traits; however, the procedure requires computationally intensive permutations for inference, so it is unclear whether the approach is scalable to large-scale resequencing efforts.

In our project, we will adapt the QTDT (Abecasis, Cardon et al. 2000) framework to overcome the issue induced by population stratification. The QTDT framework decomposes each genotype into a between-family component (sensitive to population stratification) and a within-family component (robust to population stratification). This kind of decomposition was first introduced in Fulker et al. (Fulker, Cherny et al. 1999), which mainly addressed the situation where only siblings are available. Abecasis (Abecasis, Cardon et al. 2000) extended this method by incorporating parental information into the construction of between-family components. The method to calculate within-family components (Saad and Wijsman) and between-family components (*Bij*) is very intuitive:

$$B_{ij} = \left\{ \begin{array}{l} \text{Average genotype of parents, if parental information is available} \\ \text{Average genotype of siblings, if parental information is not available} \end{array} \right\}.$$

After *Bij* is obtained, *Wij* can be obtained by subtracting the between-family component from the genotype:

$$W_{ij} = G_{ij} - B_{ij}. \qquad (6)$$

If the minor allele is considered the reference allele, a positive value of *Wij* means excess transmission of the minor allele while a negative value means excess transmission of the major allele. Following the biometric model, Abecasis et al. model the estimated value of the outcome through the following model:

$$\overset{Ù}{y}_{ij} = m + b_a g_{ij}, \qquad (7)$$

where *gij* is the genotype score of the *j*th individual in the *i*th family, and n is the population mean. They further assume that the mean trait value for individual *j* in the *i*th family takes the form:

$$E(y_{ij}) = E(m + g_{ij}a), \qquad (8)$$

where *a*, the additive genetic effect value, is the expected value of b_a. However, if the population is admixed, and families were sampled from different populations, then

$$E(y_{ij}) = E(m_l + g_{ij}a), \qquad (9)$$

and

$$E(\beta_a) = \frac{\sum_i n_i(p_i - q_i)\mu_i}{NV_g} + a, \qquad (10)$$

where μ_i is the population mean of the i^{th} population, n_i is the number of children in each family, and $N = \sum_i n_i$ is the total number of children. p_i and q_i are allele frequencies of the population from where the i^th^ family was drawn, and V_g is the variance of the genotype score. It can be seen that under population stratification, the expected value of β is different for each sub-population. However, after decomposing the genotype into the within-family component and the between-family component, Abecasis et al.(Abecasis, Cardon et al. 2000) showed that

$$E\begin{pmatrix} \beta_b \\ \beta_w \end{pmatrix} = \begin{pmatrix} \dfrac{\sum_i n_i(p_i - q_i)\mu_i}{NV_b} + a \\ a \end{pmatrix}, \qquad (11)$$

where V_b is the variance of the between-family components. From (9) we can tell that only the between-family component is influenced by the population structure, since it includes μ_i while the within-family component is free of μ_i. This shows that tests based solely on the within-family component are robust to population stratification.

1.2.4 Screening methods

One potential drawback of tests using only the within-family component is that ignoring the between-family component might lead to power loss. We borrowed ideas of Purcell et al. (Purcell, Sham et al. 2005) and Van Steen et al. (Van Steen, McQueen et al. 2005) to adopt a two-stage screening procedure. Van Steen's (Van Steen, McQueen et al. 2005) method was originally designed to overcome the multiple comparison issue in GWAS to increase power.

Their strategy is to first estimate the genetic effect for each SNP, screen and test these effects at the first stage, and at the second stage, only test on the top SNPs identified from the first stage. In order to avoid bias, data used for testing in stage one of the screening process needs to be independent of the data used to test in the second stage. However, due to the unique advantage of the QTDT framework, the between-family component and the within-family component are orthogonal to each other (Fulker, Cherny et al. 1999); hence, we do not need to recruit different individuals to test in the second stage. In our study, we test on the between-family component in the first stage, while in the second stage we test on the orthogonal within-family component. Purcell et al.'s method (Purcell, Sham et al. 2005) showed that incorporating parental information in the test could increase power. We borrowed their idea and screen on parental genotype at the first stage while testing on the offspring-based within-family component at the second stage.

1.3 Summary

As discussed above, family-based rare-variant sequencing studies have several advantages compared to their population counterparts. However, existing methods for family-based designs mostly ignore the potential bias caused by population stratification. As more and more studies adopt family designs to identify traits associated rare-variants, powerful and robust statistical methods are needed. In my dissertation research, I will focus on developing statistical methods for rare variant studies in family designs that are robust to population stratification while maintaining high power. In the following chapters, I will present my first method in Chapter 2: Flexible and Robust Methods for Rare-Variant Testing of Quantitative Traits in Trios and Nuclear Families. In this method, we integrate the QTDT framework with the kernel-machine framework, and adopted two screening methods to improve power. Our method has the advantages of analytical calculation of p-values and robustness to population stratification. This work is now published in *Genetic Epidemiology* (Jiang, Conneely et al. 2014). In Chapter 3, I will present my second project titled 'Robust Method for Rare Variant Testing of Quantitative Traits in General Pedigrees. This method preserves all the advantages of our previous method described in Chapter 2, but can be applied to large pedigrees of arbitrary size and structure; statistical methods for the analysis of large pedigrees are generally lacking in the literature. This work has now been submitted to *Statistics in Biosciences* and is in minor revision at the journal. Next, in chapter 4, I will present my third project titled 'Powerful and Robust Cross-Phenotypes Association Tests of Rare Variants in Case-Parent Trios', which takes the QTDT framework used in Chapter 2 and implements it into KDC-based testing of multiple phenotypes simultaneously to allow for robust testing of pleiotropy in family studies. Finally, in Chapter 5, I will provide a conclusion section summarizing our findings and describe future extensions and areas of research.

Chapter 2. Flexible and Robust Methods for Rare-Variant Testing of Quantitative Traits in Trios and Nuclear Families

This chapter was published in Genetic Epidemiology (DOI 10.1002/gepi.21839)

ABSTRACT

Most rare-variant association tests for complex traits are applicable only to population-based or case-control resequencing studies. There are fewer rare-variant association tests for family-based resequencing studies, which is unfortunate since pedigrees possess many attractive characteristics for such analyses. Family-based studies can be more powerful than their population-based counterparts due to increased genetic load and further enable the implementation of rare-variant association tests that, by design, are robust to confounding due to population stratification. With this in mind, we propose a rare-variant association test for quantitative traits in families; this test integrates the QTDT approach of Abecasis et al. (Abecasis, Cardon et al. 2000) into the kernel-based SNP association test KMFAM of Schifano et al. (Schifano, Epstein et al. 2012). The resulting within-family test enjoys the many benefits of the kernel framework for rare-variant association testing, including rapid evaluation of p-values and preservation of power when a region harbors rare causal variation that acts in different directions on phenotype. Additionally, by design, this within-family test is robust to confounding due to population stratification. While within-family association tests are generally less powerful than their counterparts that use all genetic information, we show that we can recover much of this power (while still ensuring robustness to population stratification) using a straightforward screening procedure. Our method accommodates covariates and allows for missing parental genotype data, and we have written software implementing the approach in R for public use.

2.1 Introduction

The emergence of next-generation sequencing technology, along with the development of the exome chip, have led many investigators to study the role of rare genetic variation in complex human traits. Rather than analyze rare variants individually, many statistical approaches for rare-variant association mapping employ grouping strategies that aggregate rare variants in a gene or region for analysis to improve power. These approaches can be broadly categorized as either burden tests that collapse grouped rare variants into a single aggregate variable that is then regressed on phenotype (Kwee, Liu et al. 2008, Madsen and Browning 2009, Morris and Zeggini 2010, Zawistowski, Gopalakrishnan et al. 2010), kernel tests that relate phenotype to rare variants in a region as a function of a variance component (SKAT, (Wu, Lee et al. 2011)), and unified tests that combine burden and kernel tests together (SKAT-O, (Lee, Emond et al. 2012)). Burden tests are preferred when a region harbors rare causal variants that all act in the same direction on phenotype (all protective or all deleterious) whereas kernel tests are optimal when a region harbors rare causal variants that act in different directions on phenotype (Wu, Lee et al. 2011).

Although these rare-variant methods generally have improved power compared to tests of individual rare variants, almost all of these tests are restricted to case-control or population-based study designs and cannot be used in family-based studies. Family-based designs have several advantages over population-based designs in that they enable the use of statistics that, by design, are robust to confounding due to population stratification. Family designs also can solve genetic problems that are hard to answer in population-based studies. For example, sequencing the parents of affected subjects can identify *de novo* mutations and also allow the study of rare homozygous genotypes, which are difficult to find in population-based designs (Do, Kathiresan et al. 2012). Families are also attractive to study because they often provide increased genetic load for a disease or trait: while carriers of a minor risk allele will be hard to sample in the general population, they are more likely to be found in families of probands (Zollner 2012). Finally, family studies allow the study of the segregation pattern of complex disease (Ott, Kamatani et al. 2011). Because of these appealing features and the fact that there are many familial samples from past linkage studies, family-based resequencing studies are gaining in popularity. Several recent studies have identified disease-associated rare variants through family-based designs, including rare variants associated with multiple sclerosis (Ramagopalan, Dyment et al. 2011), simplex autism (Krumm, O'Roak et al. 2013), dilated cardiomyopathy (Norton, Li et al. 2011), and Alzheimer's disease (Cruchaga, Haller et al. 2012).

Recently, a few methods have been proposed for rare-variant association testing in families. Schaid et al. (Schaid, McDonnell et al. 2013) developed a method for complex traits that accounts for relatedness among study subjects. Their method took a retrospective view of the sample, which assumes that the outcome is fixed while the genotype is random, and is particularly appealing for the analysis of datasets that are collected under non-random ascertainment (such as those collected for linkage studies). Chen et al. (Chen, Meigs et al. 2013) developed a rare-variant test for quantitative traits in families by extending kernel-machine methods (Kwee, Liu et al. 2008, Wu, Lee et al. 2011) to pedigree analysis by inserting a random familial effect due to shared polygenes within the modeling framework; a similar idea was employed by Schifano et al. (Schifano, Epstein et al. 2012) and Oualkacha et al (Oualkacha, Dastani et al. 2013). Jiang et al. (Jiang and McPeek 2014) adopted the similar strategy and extended the SKAT-O (Lee, Emond et al. 2012) method to family studies of quantitative traits. Although the methods of both groups adjust for kinship in family studies, they do not consider potential bias caused by population stratification. Population stratification can lead to substantially inflated false-positive rates in sequencing studies of rare variants (Epstein, Duncan et al. 2012, Jiang, Epstein et al. 2013, Liu, Nicolae et al. 2013), and standard GWAS approaches to correct for such stratification (such as principal components or EMMAX (Kang, Sul et al. 2010)) may not be effective when applied to rare variants (Mathieson and McVean 2012). Therefore, a rare-variant association test that maintains validity in the presence of such stratification is needed. Ionita-Laza et al. (Ionita-Laza, Lee et al. 2013) proposed such a method based on the family-based association test (FBAT) framework. Although this method is robust to population stratification, it ignores between-family information that could perhaps be exploited to boost power. Fang et al. (Fang, Sha et al. 2012, Fang, Zhang et al. 2013) used between-family information for this purpose in an adaptive rare-variant association test for quantitative traits; however, the procedure requires computationally intensive permutations for inference, so it is unclear whether the approach is scalable to large-scale resequencing efforts.

In this paper, we propose a novel two-stage method for rare-variant analysis of quantitative traits in trios and nuclear families. The approach is based on the QTDT (quantitative transmission disequilibrium test) framework of Abecasis et al. (Abecasis, Cardon et al. 2000) for SNP association mapping. The QTDT framework decomposes the observed individual genotypes into between-family and within-family components. The within-family component is robust to population stratification, while the between-family component is sensitive to the phenomenon. In this paper, we calculate the within-family component for each rare variant in a region, and then integrate these components within the kernel procedure KMFAM of Schifano et al. (Schifano, Epstein et al. 2012), which was previously developed for SNP-set association testing of quantitative traits in families. Specifically, within KMFAM, we create a kernel matrix based on the within-family component, and then use this kernel matrix to test for association with phenotype using a modified score statistic. By using the within-family component only, our rare-variant association test for quantitative traits is robust to confounding due to population stratification. Also, the approach calculates p-values analytically rather than via resampling and is thus scalable to exome sequencing and whole-genome resequencing studies. Because the approach relies on a kernel framework, it also preserves power when a region contains a mixture of trait-increasing and trait-decreasing variation. The approach also allows for covariates and, for nuclear families, can be implemented when phenotype and genotype data on parents are missing, so it can be applied in the study of quantitative traits related to late-onset diseases.

A potential drawback of using only within-family information for analysis is that power is reduced by ignoring the (sensitive) between-family information within the analysis (Ionita-Laza, Lee et al. 2013). However, borrowing ideas from Purcell et al. (Purcell, Sham et al. 2005) and Van Steen et al. (Van Steen, McQueen et al. 2005), we propose using between-family information as a screening tool to identify the most interesting regions (based on the magnitude of the p-value for the region) that merit further investigation. We then apply our within-family test to only these top regions, thereby reducing the multiple-testing burden (compared to within-family testing of all regions) and potentially gaining power. We note that the first stage of the analysis (using the between-family information) is independent of the second stage (which uses orthogonal within-family information). We also note that, by using within-family information in the second stage, our approach is still robust to confounding due to population stratification.

In subsequent sections, we first describe the KMFAM procedure and then, for rare variants, discuss how we integrate the QTDT framework into the model to make the method robust to population stratification. We next describe our screening procedure to improve power.

We then evaluate our approaches using simulated sequence data in trios and nuclear families and show how screening can improve power of within-family testing while maintaining an appropriate type I error rate, even under population stratification. Finally, we summarize our method and discuss potential extensions.

Materials and Methods

2.2.1 Notation and KMFAM Model: We initially present the KMFAM model of Schifano et al. ((Schifano, Epstein et al. 2012) (also used by Chen et al. (Chen, Meigs et al. 2013)), and then show how to modify the framework to develop a within-family association test of rare variation for quantitative traits. As in KMFAM, we assume a sample of N nuclear families that are genotyped for s rare-variants in a gene or region of interest. Let Y_{ij} denote the quantitative outcome for the j^{th} individual in the i^{th} family, where i=1,2,3…N and j=1,2,..n_i. We define X_{ij} as a c×1 vector that represents the covariates for the j^{th} individual in the i^{th} family and further define G_{ij} as an s×1 vector that represents the genotypes of the s rare variants for each subject (where each rare-variant genotype is coded as the number of copies of the rare allele the subject possesses at each site). We assume that the outcome, Y_{ij}, follows a multivariate normal distribution with mean and variance defined through the model:

$$Y_{ij} = X_{ij}^T a + G_{ij}^T b + f_{ij} + e_{ij}, \qquad (1)$$

where a is a c×1 vector of coefficients for X_{ij} and b is a sx1 vector of coefficients for G_{ij}. While we assume the coefficients in a are fixed effects, we instead assume the coefficients for the genotype effects b are random and follow an arbitrary distribution with variance t. With this assumption, we can test for association between rare variants and phenotype by considering the hypothesis $t = 0$ rather than an *s* degree of freedom fixed-effects test: $b_1 = b_2 = b_3 = ... = b_s =$ 0, which will have low power.

To complete the formulation of model (1) for pedigree data, we let f_{ij} denote the random effect to account for within-family correlation due to shared polygenes. We assume the effect within a family follows a multivariate normal distribution: $f_i \sim MVN(0, 2F_i s_{pg}^2)$, where F_i is the kinship matrix for family *i* and $s_{pg}^2$is the variance due to the effect of polygenes. We also define $e_{ij} \sim N(0, s_e^2)$as the random error term. From model (1), we calculate the variance of outcome as

$$V = Var(Y) = tK + s_{pg}^2 2F_i + s_e^2 I \ , \qquad (2)$$

where *K*=***GIG****^T^* is the kernel matrix, and ***G*** is a matrix composed of the vectors G_{ij} such that each row is G_{ij}^T for a single individual. Note that here we use a linear kernel, but if previous information is available for the rare variants in the gene, the use of other kernels, such as a linear weighted kernel, can increase power (Wu, Lee et al.); in this case *I* can be replaced with a weighting matrix *Z*, where elements in *Z* represent the weight. There are several methods to specify the weight, based on the belief of the variant's contribution to the outcome. One common method is to calculate weight as a function of the minor allele frequency (MAF); Wu et al. (Wu, Lee et al.) considered such a weight that modeled MAF using a Beta distribution, but other weights are possible, as well.

To test whether the rare variants in the gene are associated with the outcome, we construct a variance component score test derived from model (1) (Lin 1997, Zhang and Lin 2003). The null hypothesis is H$_0$:$t = 0$, and the test statistic takes the form:

$$Q = \frac{1}{2}(Y - X\overset{Ù}{a}_0)\overset{Ù}{V}_0^{-1} K \overset{Ù}{V}_0^{-1}(Y - X\overset{Ù}{a}_0), \qquad (3)$$

where all parameters are estimated under the null hypothesis. We define $\hat{V}_0$ and $\hat{\alpha}_0$ as the estimates of V in (1) and α under the null. Further, we define a projection matrix $P = \hat{V}_0^{-1} - \hat{V}_0^{-1} X(X^T \hat{V}_0^{-1} X)^{-1} X^T \hat{V}_0^{-1}$, such that $PV_0P = P$. Thus, under the null, we have

$$Q = \frac{1}{2} Y^T PKPY = \sum_{i=1}^{N} \lambda_i \chi_{1i}^2, \qquad (4)$$

where λ_i are eigenvalues of $\frac{1}{2} D \hat{V}_0^{-1/2} K \hat{V}_0^{-1/2} D$, here $D = I - \hat{V}_0^{-1/2} X(X^T \hat{V}_0^{-1} X)^{-1} X^T \hat{V}_0^{-1/2}$. As χ_{1i}^2 are independently and identically distributed random variables, *Q* is distributed as an asymptotic mixture of chi-square distributions, and the p-values can be calculated using the Davies method (Davies 1980).

<u>*2.2.2 Robust Rare-Variant Association Test:*</u> One issue with the KMFAM framework described above is that the resulting score tests from model (1) are sensitive to population stratification. To resolve this issue, we integrate the QTDT (Abecasis, Cardon et al. 2000) framework into our model. The QTDT framework decomposes the observed genotype G_{ij} into a between-family component (which we denote by B_{ij}) and an orthogonal within-family component (which we denote by W_{ij}). The between-family component takes the following value:

$$B_{ij} = \left\{ \begin{matrix} \text{Average genotype of parents, if parental information is available} \\ \text{Average genotype of siblings, if parental information is not available} \end{matrix} \right\}.$$

Once we obtain the between-family component, we then construct the within-family component, W_{ij}, by subtracting the between-family component from the observed genotype such that $W_{ij}=G_{ij}-B_{ij}$.

By design, association analyses of complex traits that base inference on the within-family component W_{ij}, are robust to population stratification. Based on this observation, we can construct a robust rare-variant association test for trios and nuclear families by replacing the observed genotypes $\boldsymbol{G}$ in the kernel matrix $\boldsymbol{K}$ described in (2) with their corresponding within-family components $\boldsymbol{W}$. We then construct the score statistic Q in (3) as before to derive our robust family-based association test.

2.2.3 Screening Procedure: Although the QTDT framework ensures the robustness of our proposed score test to potential confounding due to population stratification, the discarding of between-family information when confounding due to population stratification is not an issue can lead to sizable power loss compared to use of the observed genotype. In attempts to restore the power of our within-family association test to levels anticipated when using observed-genotype information, we suggest a two-stage screening approach that uses both the within- and between-family rare-variant information. In the first stage, we use between-family information to screen and identify the top regions for follow up. If parental phenotype and genotype information are available, we carry out the first stage by performing the SKAT (Wu, Lee et al. 2011) test on parents only, and then select a subset of regions for follow-up investigation based on smallest p-values. If parental information is unavailable, we instead conduct the first-stage screening by applying KMFAM to the outcomes and between-family components of the offspring. In the second stage, we construct the robust test (using the within-family components calculated for the offspring) only on those top regions selected from the first stage. By only testing a reduced number of regions in the second stage using the within-family component, we reduce the number of robust tests that are conducted thereby reducing the multiple-testing burden and increasing power. As discussed in Abecasis et al. (Abecasis, Cardon et al.), the between-family and within-family components are orthogonal to each other, such that the first-stage and second-stage tests are independent.

2.2.4 Type I Error Simulations: We evaluated the type I error and power of our approach using simulated sequencing data. We used *cosi* (Schaffner, Foo et al. 2005) to simulate sequence data for a pool of 5000 European and 5000 African haplotypes, each of length 30 kb. Rare variants were defined as variants with a minor-allele frequency less than 3% in the region. To simulate family data, we randomly paired subjects within each population and simulated offspring by sampling one haplotype from each parent. When considering nuclear families with 2 or more offspring, we performed simulations for the situation where all parental information is available, as well as where 20-100% parental information is missing.

Using this concept, we first performed type 1 error rate simulations to verify that our method is robust to population stratification. We simulated the outcome through the null model:

$$Y_{ij} = gI_{African,\, ij} + f_{ij} + e_{ij}, \qquad (5)$$

where g is the mean trait difference between European and African subjects, $I_{African,\, ij}$ is an indicator variable that is 1 if the subject is African and 0 otherwise, and all other terms are the same as defined in model (1). We specified f_{ij} and e_{ij} such that the overall trait heritability was 0.35. To induce confounding due to population stratification in our simulations, we first assumed our sample consisted of a mixture of European and African families, with the percentage of European families ranging from 25% to 75%. We then assumed a value of g in model (5) that ranged from 0 (no confounding due to population stratification) to 3 (extreme confounding due to population stratification).

2.2.5 Power Simulations: To estimate power, we simulated a region of 300 kb, divided into 10 non-overlapping regions of 30 kb each, and selected one region at random as causal (the other 9 regions are assumed to be independent of outcome). To generate trait data for each subject based on the causal region, we used the idea of Wu et al. (Wu, Lee et al. 2011) and assumed a certain percentage (5% or 15%) of rare variants (defined as variants with a minor allele frequency less than 3%) in the region influenced the outcome, with the effect size of a causal variant defined as $b = c' \; |\log_{10} MAF|$, where we varied the constant *c* among values between 0.4 and 0.6. We then included these effects due to rare variants within model (5) to simulate the outcome. To keep power at a reasonable range for the 300kb region, we fixed g at 0.25 for power simulations under stratification. As with the null simulations, we assumed the trait heritability was 0.35.

2.3 Results

2.3.1 Type I Error: We first performed type I error rate simulations on parent-offspring trios to demonstrate that population stratification can lead to spurious rare-variant association with quantitative traits in families. Figure 2.1 presents type I error results for two methods: our robust rare-variant approach that uses within-family information from the offspring only and a SKAT test of rare-variant association that uses the observed offspring genotype (constituting both the within- and between-family components). For these simulations, simulated datasets consisted of 500 trios where 50% are of European descent and the remaining 50% are of African descent. When the mean trait difference between European and African populations is 0 (such that there is no confounding due to population stratification), both the within-family test and observed-genotype test had appropriate type I error. However, when we induced confounding due to population stratification by assuming a non-zero mean trait difference between Africans and Europeans, we found the standard SKAT test using the observed genotype had inflated type I error. Our robust rare-variant association test, in contrast, maintained the proper type I error rate under confounding.

We next performed another set of type I error simulations, where we assumed datasets consisting of 500 nuclear families each with two children. We varied the proportion of nuclear families that were of European origin between 25% and 75% and assumed the mean trait difference between African and European samples to be 2 (thereby inducing confounding due to population stratification). We further assumed the proportion of nuclear families within each dataset that was missing parental genotype information ranged from 0% to 100%. In our first set of simulations (shown in Figure 2.2), we studied the type I error rates of methods assuming examination of the 30-kb region in its entirety. We compared the type I error rates using the observed genotype information in the offspring only (which corresponds to the test of Chen et al. (Chen, Meigs et al. 2013)), as well as using our robust rare-variant association test that relies only on the within-family information in the offspring. Our results indicated that rare-variant association tests using observed genotype information led to considerable inflation in type I error rates across different simulation models, whereas our robust within-family association test remained valid in all situations. The validity of the robust rare-variant association test was confirmed both when parental genotype information was available on all participants, as well as when such genotype information was completely absent in the dataset. Thus, for late-onset diseases in which parental information might not be available, our method is still robust to population stratification.

We performed a final set of type I error simulations for nuclear families of size two under our proposed screening scheme where, in this instance, we split the 300-kb region into 10 non-overlapping regions, each of size 30 kb. Using between-family information, we identified a subset of regions for follow up (based on p-value) that we then investigated further using the within-family component. Our results are shown in Figure 2.3. Overall, our results show that our screening procedure (conducted using either parental information or between-family information in siblings, if parents are not available) preserved type I error across models, with differing missing parental information as well as different proportions of regions that were then followed up using within-family information. These results demonstrate that our screening procedure maintains appropriate type I error, even when there is confounding due to population stratification, due to the fact that the between-family component and within-family component of the offspring genotype are orthogonal to one another.

2.3.2 Power: In the previous section, we showed that our robust rare-variant association tests that uses the within-family component remains valid in the presence of population stratification. We next studied the power of our proposed robust test to detect association with a trait under various trait-influencing models. We assumed either 5% or 15% of rare variants in a region were causal and assumed the effect size of such causal variants was $b = c' \; |\log_{10} MAF|$, where c ranged from 0.4 to 0.6. We first compared the power of our robust within-family association test to the standard observed-genotype test considered by Chen et al. and Schifano et al. under models with no population stratification (to ensure the power of the observed-genotype test was valid). We generated sequence and trait data on 500 nuclear families each with two offspring. We first analyzed the observed rare-variant genotypes in the family using the kernel test of Chen et al., and then repeated the analysis using our robust within-family association test. As shown in Figure 2.4, the power of the kernel test using observed genotype information (shown as black bars) is, as expected, more powerful than the same test using within-family information alone (shown in gray bars) across different simulation models. In attempts to see whether we could restore some power to the robust test, we then applied our screening procedure to these simulated datasets using between-family information. For each dataset, we tested the between-family components of each of the 10 regions, and then subsequently considered only the top 10%, 20%, 30%, or 40% (based on minimum p-value) of these regions using our within-family test. The results show that, when screening is performed using parental genotype and trait information, our screening procedure restores power to levels similar to those using the observed-genotype information (see top panels of Figure 2.4). If screening is instead performed using between-family information, the robust within-family association test also shows a power increase, although it is not as notable as using parental information (see bottom panels of Figure 2.4). Thus, it appears that our initial screening step improves the power of the within-family association test, while preserving appropriate type I error under the null.

While we obtained our results in Figure 2.4 under simulation models that assumed no confounding due to population stratification, we also observed simulated trends in simulation models that were generated with confounding due to population stratification. Figure 2.5 presents power results under confounding due to population stratification that assumed a mean trait difference between African and European samples. As the observed-genotype test under confounding is not valid, here we report the empirical adjusted power (black bars). To get the empirical adjusted power, we first simulated under the null distribution at the present of population stratification and get the confounded empirical distribution. We then adjusted the observed genotype’s power based on this empirical distribution. The remaining bars denote the power of the robust within-family association test, along with variations that screen using parental or between-family information. The results show that screening can improve power of the robust rare-variant test, particularly as the percentage of causal variants and the magnitudes of their effect increase. The results in Figure 2.5 were for simulated datasets consisting of nuclear families with two offspring each; we saw similar trends when analyzing parent-offspring trios, as well (see Supplemental Figure 2.1).

2.4 Discussion

In this chapter, we proposed a kernel method for analyzing rare-variant sequencing studies in trios and nuclear families that is robust to confounding due to population stratification. We also introduced a screening procedure using parental or between-family information to improve the power of this robust test and showed that this procedure can increase power to levels near those of the observed-genotype test when confounding due to stratification is not an issue. In addition to robustness, our approach has many other practical features. The method easily allows for covariates and permits rapid calculation of p-values using analytic procedures. We have implemented our procedure in R software, which is available from our website (see Web Resources). Our approach is computationally efficient, as the analysis of a 30-kb region for 500 nuclear families each of size two takes on average 53.08 seconds on a 768 processor running Linux OS with 2.6 gigahertz of RAM. Based on the computational speed, we believe the approach can be scaled reasonably to whole-exome or whole-genome resequencing studies on a multi-node cluster.

Family-based genetic studies of complex traits occasionally have information available from additional unrelated singletons. While we cannot use these individuals within our robust within-family association test of rare variation, the information from such singletons can be used in our screening step (treating them in the same way as the parental information) to identify the most interesting regions for follow up using the robust test. Such information could be helpful in screening and should not affect the validity of the second-stage robust test, even if there is confounding due to population stratification and/or coverage differences between the family and unrelated arms of the study.

Figure 2.1

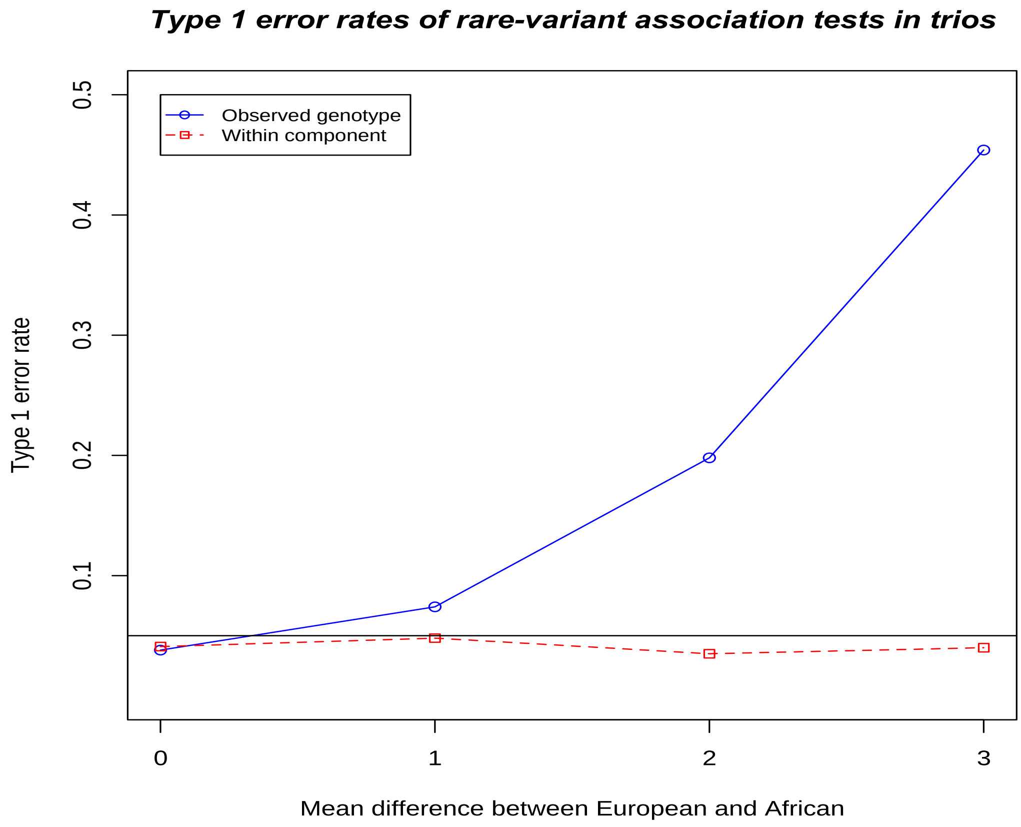


Figure 2.1: Empirical type 1 error rates of rare-variant association tests applied to 30-kb sequenced regions in parent-offspring trios. Simulated datasets consisted of 500 parent-offspring trios (50% of European ancestry, 50% of African ancestry). The mean trait difference between European and African subjects varies from 0 (no stratification) to 3 (extreme stratification). Total trait heritability is 0.35. We analyzed each simulated trio dataset twice: once using SKAT to analyze the observed offspring genotypes (“Observed genotype,” blue line) and once using our proposed kernel test that used only the within-family component of the observed offspring genotypes (“Within component,” red line). Each result is based on 1000 replicates.

Figure 2.2

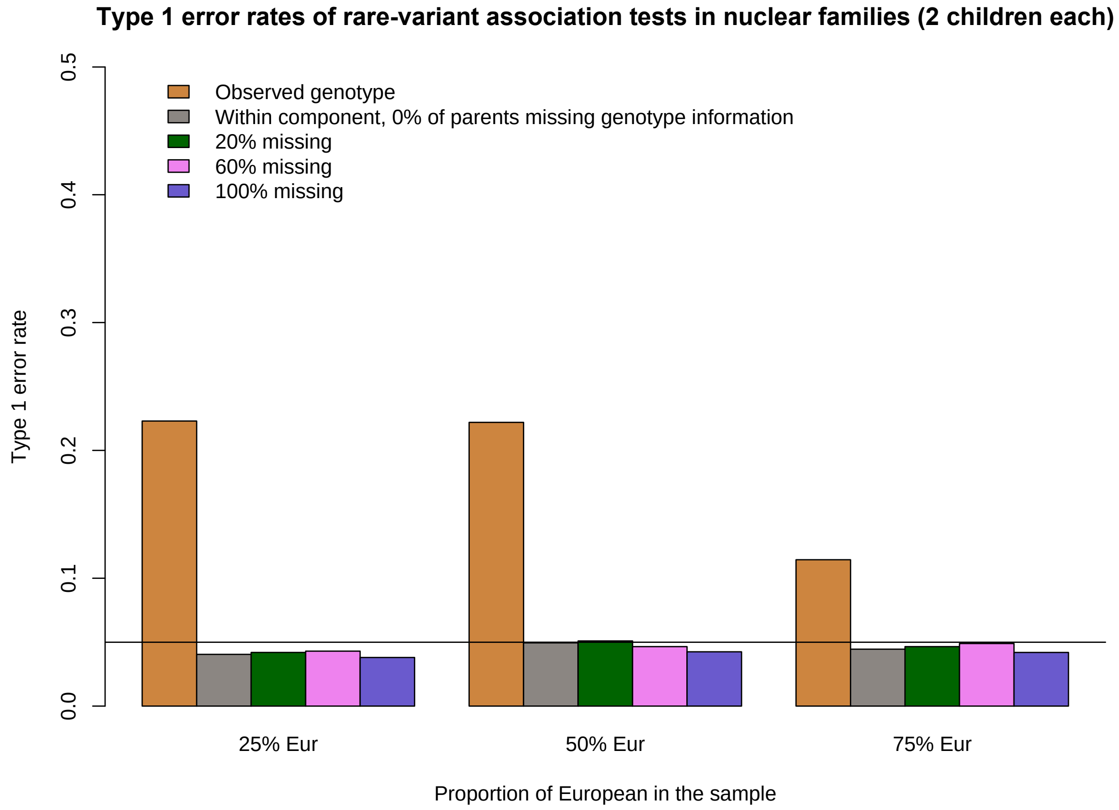


Figure 2.2: Empirical type 1 error rates of rare-variant association tests applied to 30-kb sequenced regions for nuclear families with 2 children each (total heritability is 0.35). Simulated datasets consisted of 500 nuclear familes each with 2 children. Percentage of European varies from 25% to 75%. Percentage of missing parents varies from 0% to 100%. The mean trait difference between European and African subjects is 2. For each simulated dataset, we used SKAT to analyze the observed offspring genotypes ("Observed genotype," brown bars) and used our proposed kernel test that used only the within-family component of observed offspring. For within-family results, we present findings assuming percentage of missing parents was 0%, 20%, 60%, and 100%. Each result is based on 1000 replicates.

Figure 2.3

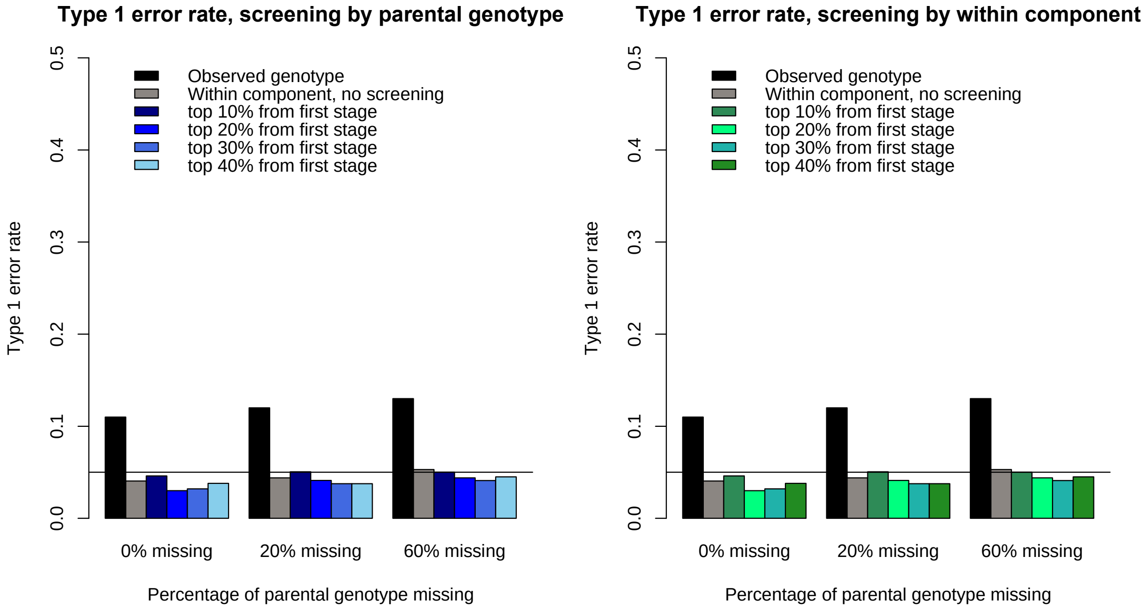


Figure 2.3: Empirical type 1 error rates of rare-variant association tests applied to ten 30-kb sequenced regions for nuclear families with 2 children each (total heritability is 0.35). Simulated datasets consisted of 500 nuclear families each with 2 children. The mean trait difference between European and African subjects is 0.25. For each simulated dataset, we first used KMFAM to analyze the observed offspring genotypes ("Observed genotype," black bars); we then used our proposed kernel test to analyze the within-family component of offspring without screening, and then applied screening procedures. We applied two screening processes: screening by parental information (blue bars) and screening by the between-family component (green bars). Left: screen by parental genotype. Right: screen by within-family component. Top 10% to 40% of regions with smallest p-value were selected through the screening process and analyzed in the second stage. Each result is based on 1000 replicates.

Figure 2.4

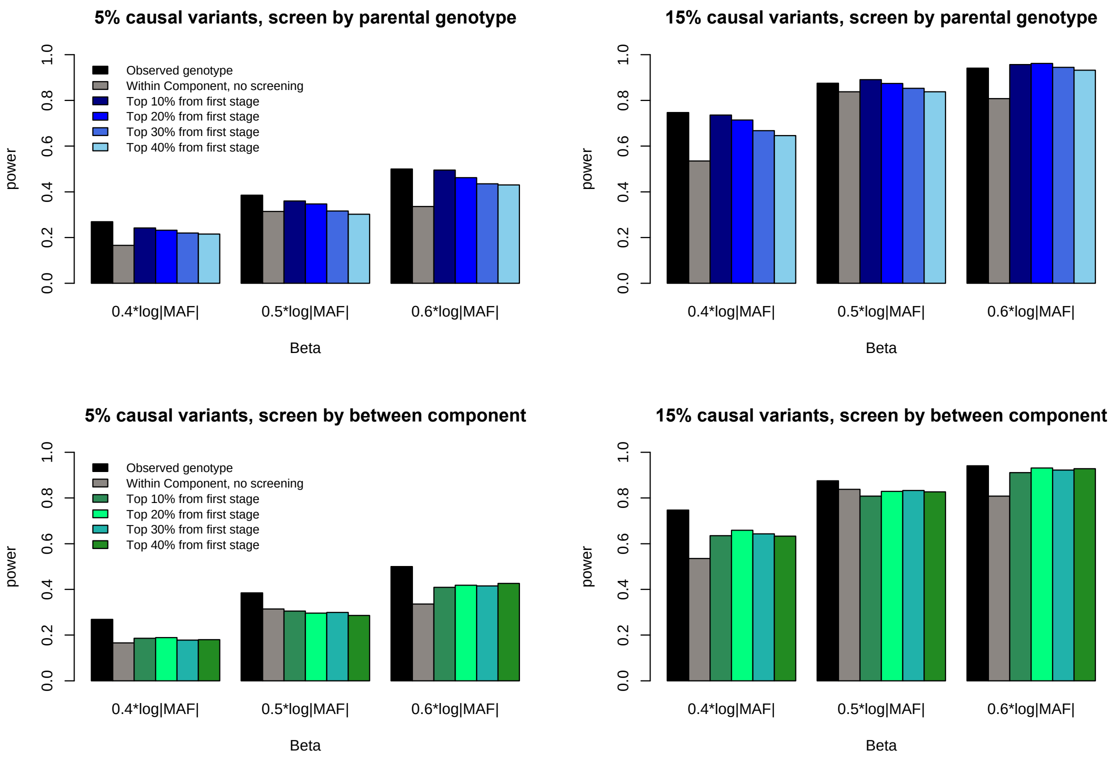


Figure 2.4: Empirical power of rare-variant association tests applied to ten 30-kb sequenced regions for nuclear families without stratification. Simulated datasets consisted of 500 European families each with 2 children. Three effect sizes were used: $0.4´\ |\log_{10} MAF|$, $0.5´\ |\log_{10} MAF|$, and $0.6´\ |\log_{10} MAF|$. As in Figure 2.3, for each dataset we used KMFAM to test the observed genotype; then we used our method to test the within-family component without screening, and then applied two screening methods. Top panel: screen by parental genotype. Bottom panel: screen by between-family component. Each result is based on 1000 replicates.

Figure 2.5

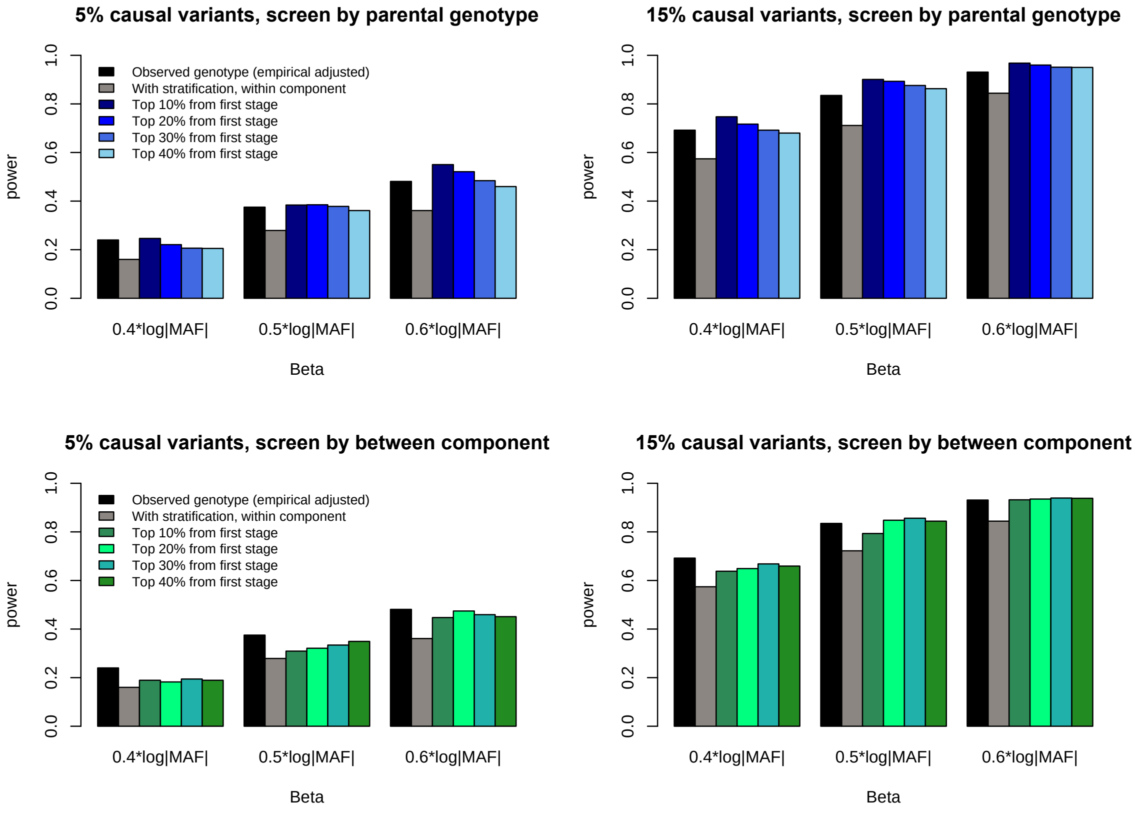
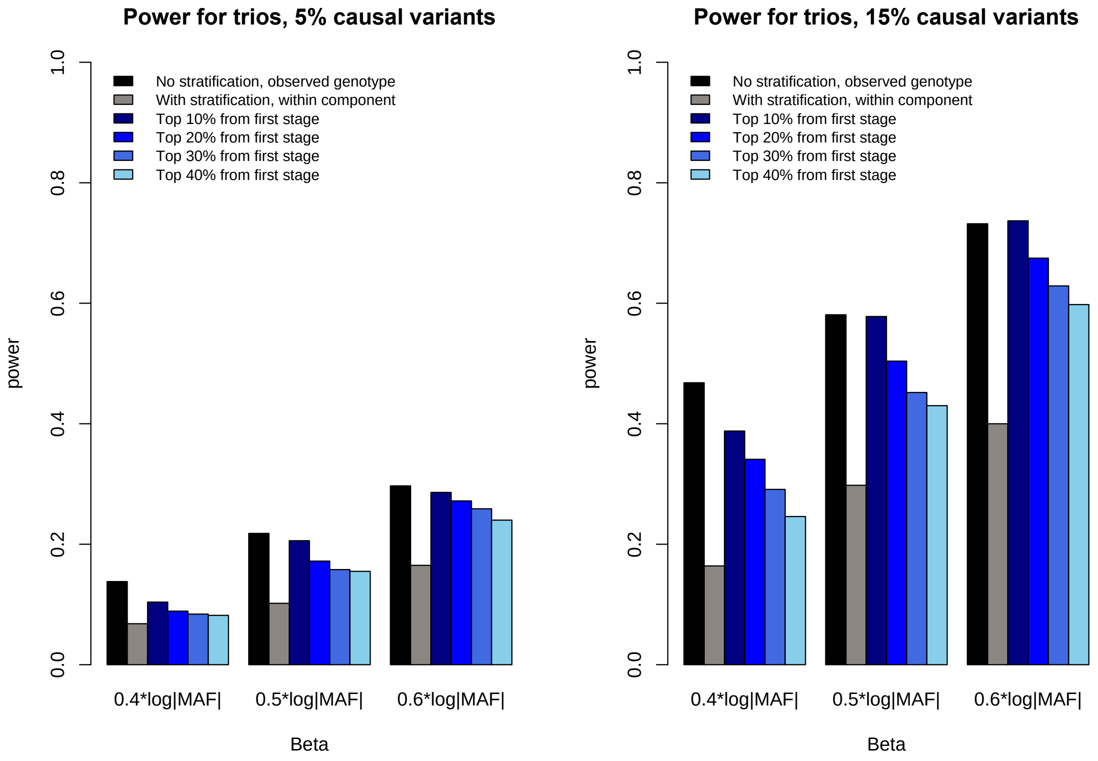


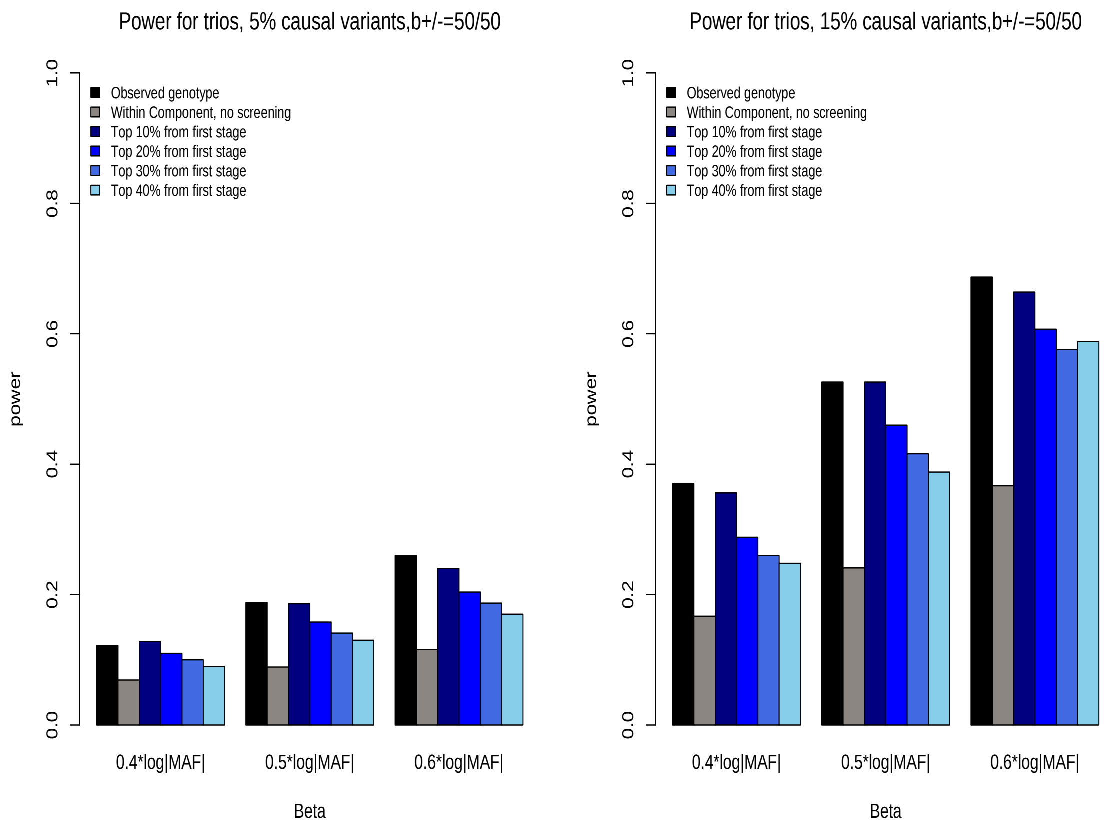
Figure 2.5: Empirical power of rare-variant association tests applied to ten 30-kb sequenced regions for nuclear families with/without stratification. Black bars are adjusted empirical power. Other simulations were performed under population structure such that 25% of families are European, and the mean trait difference between European and African subjects is 0.25. Three effect sizes were used: $0.4\acute{}\ |\log_{10} MAF|$, $0.5\acute{}\ |\log_{10} MAF|$, and $0.6\acute{}\ |\log_{10} MAF|$. Top panel: screen by parental genotype. Bottom panel: screen by between-family component. Each result is based on 1000 replicates.

Supplementary Figure .2.1



Supplementary Figure 2.1: Empirical power of rare-variant association tests applied to ten 30-kb sequenced regions for trios with/without stratification. All causal variants have positive effect on the trait value. Black bars are baseline power (no stratification). Other simulations were performed under population structure such that 25% of families are European, and the mean trait difference between European and African subjects is 0.25. Three effect sizes were used: 0.4´ $|\log_{10} MAF|$, 0.5´ $|\log_{10} MAF|$, and 0.6´ $|\log_{10} MAF|$. Left panel: 5% of rare variants in the region are causal. Bottom panel: 15% of rare variants in the region are causal. Each result is based on 1000 replicates.

Supplementary Figure 2.2



Supplementary Figure 2.2: Empirical power of rare-variant association tests applied to ten 30-kb sequenced regions for trios with/without stratification. 50% causal variants have positive effect, 50% causal variants have negative effect on the trait value. Black bars are baseline power (no stratification). Other simulations were performed under population structure such that 25% of families are European, and the mean trait difference between European and African subjects is 0.25. Three effect sizes were used: $0.4´\ |\log_{10} MAF|$, $0.5´\ |\log_{10} MAF|$, and $0.6´\ |\log_{10} MAF|$. Left panel: 5% of rare variants in the region are causal. Bottom panel: 15% of rare variants in the region are causal. Each result is based on 1000 replicates.

Chapter 3. Robust Rare-Variant Association Tests For Quantitative Traits in General Pedigrees

This chapter has been submitted to *Statistics in Biosciences* and is in minor revision

Abstract

Next generation sequencing technology has propelled the development of statistical methods to identify rare polygenetic variation associated with complex traits. The majority of these statistical methods are designed for case-control or population-based studies, with few methods that are applicable to family-based studies. Moreover, existing methods for family-based studies mainly focus on trios or nuclear families while 2nd or higher degree relatives are ignored. To fill this gap, we propose a method for rare-variant analysis in large pedigree studies that can utilize information from all available relatives. Our approach is based on a kernel-machine regression (KMR) framework, which has the advantages of high power, as well as fast and easy calculation of p-values using the asymptotic distribution. Our method is also robust to population stratification due to integration of a QTDT framework (Abecasis, et al. 2000b) with the KMR framework. In our method, we first calculate the expected genotype (between-family component) of a non-founder using all founders' information and then calculate the deviates (within-family component) of observed genotype from the expectation, where the deviates are robust to population stratification by design. The test statistic, which is constructed using within-family component, is thus robust to population stratification. We illustrate and evaluate our method using simulated data and sequence data from Genetic Analysis Workshop 18 (GAW18).

3.1. Introduction

Next-generation sequencing (NGS) studies of complex human traits and diseases are becoming commonplace for investigating the role of rare polymorphic variation in such phenotypes. Many analytic methods have been developed for the analysis of such rare variants with a particular emphasis on techniques that first aggregate information on rare variants within a gene of interest and then contrast this aggregated genetic information with the phenotypic outcome. The majority of such aggregation-based methods (Kwee, Liu et al. 2008, Madsen and Browning 2009, Morris and Zeggini 2010, Zawistowski, Gopalakrishnan et al. 2010, Wu, Lee et al. 2011, Lee, Wu et al. 2012) focus on population-based designs or case-control designs. However, family-based study designs are gaining traction in NGS projects since they provide inherent benefits over the traditional population-based designs. In particular, families ascertained based on multiple relatives with a particular phenotype tend to enrich the sample for rare causal variants compared to a general population, thereby making such variants easier to detect (Zöllner 2012).

The appeal of family-based NGS studies has lead to the development of a few analytic methods tailored for rare-variant analysis in such designs. Such methods (Chen, Meigs et al. 2013, Schaid, McDonnell et al. 2013, Jiang and McPeek 2014, Jiang, Conneely et al. 2014) generally apply a modeling framework that accounts for the relatedness of familial samples through appropriate modeling of kinship. However, such methods do not take into account the potential bias of findings due to population stratification. Population stratification is the presence of systematic differences between sub-populations both in the allele frequencies of the rare variants under study as well as in the distribution of phenotype. Failure to model these differences will lead to inflated false positive rate and decreased power to detect real associations. For rare variants, the issue of population stratification is more severe than for common variants, as rare variants are more likely to be young mutations which are more population specific (Gravel, Henn et al. 2011). It has been shown that inclusion of self-reported ethnicity as a covariate is not sufficient to adjust for population stratification (Serre, Montpetit et al. 2008). Similarly, standard methods to adjust for population stratification for common variants may not be as effective an adjustment for rare variants. In particular, genomic control can lead to very conservative results for rare variants (Jiang, Epstein et al. 2013). Although principal components works well for spatially distinctive populations, the procedure fails for spatially non-distinctive populations (Mathieson and McVean 2012).

With these concerns in mind, Jiang et al. (2014) developed a rare-variant association test for quantitative traits in parent-child trios and nuclear families that, by design, was robust to population stratification. The method was motivated by the QTDT framework (Abecasis, et al. 2000a), which showed that the observed genotype of a familial subject could be partitioned into orthogonal between-family and within-family components. The between-family component can be defined as the expected value of the subject's genotype within the family and can be constructed as the average of the parents' genotype or the average of the siblings' genotype. The within-family component is the deviation of the observed genotype from the between-family component. While the between-family component is sensitive to population stratification, the within-family component is robust to stratification since its based on a family-specific deviation. Utilizing a kernel-machine regression (KMR) framework for multi-marker analysis of familial quantitative phenotypes (Schifano, et al. 2012, Chen, et al. 2013), Jiang et al. (2014) created a robust rare-variant test by replacing observed sample genotypes in the standard KMR with their corresponding within-family genotypic components. Simulation results demonstrated the approach yielded appropriate type-I error even when strong confounding existed within the sample. As with other KMR approaches, the Jiang et al. (2014) approach derived p-values analytically using Davies' (1980) method, thereby allowing easy application to large scale sequencing studies.

The work of Jiang et al. (2014), like many other existing methods, can only be applied to parent-child trios and nuclear families. Robust and powerful methods for extended pedigrees that include 2nd or 3rd degree relatives like grandparent/grandchild and first cousins are lacking in the literature. Large pedigrees have unique features that make them ideal for mapping traits associated with rare variants. Compared to nuclear families or trios, rare variants are further enriched in large pedigrees (Wijsman 2012). It has been shown that large pedigree studies have increased power compared to smaller families with the same total number of samples, especially for rare-variant sequencing data (Wijsman and Amos 1997, Simpson, Justice et al. 2011, Wilson and Ziegler 2011). In addition to improved power, analysis of large pedigrees can provide evidence for both co-segregation and association, while population based studies can only provide evidence for association (Laird and Lange 2006, Wijsman 2012, Ott, Wang et al. 2015). Further, the study of large pedigrees provides a cost-effective strategy for rare-variant analysis as it enables *in silico* imputation of rare-variant genotypes in non-sequenced subjects using information from sequenced relatives coupled to knowledge of inheritance flow (Wijsman 2012, Cheung, Blue et al. 2014). With a large pedigree-based study design, researchers can also combine sequencing-based association studies with linkage analyses (Ott, Wang et al. 2015). Recent research has identified rare variants associated with several diseases or traits like hyperkalemic hypertension (Louis-Dit-Picard, Barc et al. 2012), spinocerebellar ataxias (Wang, Yang et al. 2010), hypolipidemia (Musunuru, Pirruccello et al. 2010), and lithium-responsive bipolar disorder (Cruceanu, Ambalavanan et al. 2013) by combining association and linkage approaches.

In this paper, we expand on the work of Jiang et al. (2014) to allow robust rare-variant analysis of quantitative traits within general pedigrees of arbitrary size and structure. To do so, we employ a modified QTDT framework for extended pedigrees developed by Abecasis et al. (2000b) that uses information from all genotyped family members to construct a more informative between-family genotypic component. We then derive the within-family component for each genotype and integrate this information within the KMR framework of Schifano et al. (2012) to obtain a rare-variant test that is robust to population stratification. In the following sections, we will first introduce our study setting, followed by how we use the QTDT framework to decompose genotype information to obtain a robust within-family component. We then show how to integrate this information within a KMR framework to yield our robust test. We will also describe how we can improve the power of our robust test by pre-screening potential trait-influencing genes using genotype and phenotypic information from founders across families. Such founder information is orthogonal to the within-family information used in our proposed test. We then evaluate our method using both simulation studies and sequencing data from a study of systolic and diastolic blood pressure (SBP and DBP) provided by the Genetic Analysis Workshop 18 (GAW18).

2. Materials and Methods

3.2.1 Study Design and Notation: We assume a family-based study consisting of *N* families, where each family consists of a large pedigree. While we use Figure 3.1 as an example here to show the structure of the large pedigree, our method can be applied to any family structure and can accommodate any family size. Suppose there are *s* rare variants in a gene of interest, and let $\boldsymbol{G}_{ij}$, a $s \times 1$ vector, represent the genotypes of the *s* rare variants for the j^{th} (j=1,2…,n_i) individual in the i^{th} (i=1,2…N) family. We assume an additive model, and let components in $\boldsymbol{G}_{ij}$ take the value of 0, 1, 2, indicating the number of copies of minor alleles at each site. If an individual is not genotyped, then we leave $\boldsymbol{G}_{ij}$ undefined. Let $\boldsymbol{X}_{ij}$, a $c \times 1$ vector, denote the covariates, and denote Y_{ij} as the value of the quantitative outcome for the j^{th} individual in the i^{th} family. For non-founders (defined as individuals with ancestors included in the pedigree, e.g. individuals 5,6,7,8,9,10 in Figure 3.1), let M_{ij} and F_{ij} be the index of mother and father of j^{th} individual in the i^{th} family respectively. For founders (defined as individuals with no ancestors in the pedigree, e.g. individuals 1,2,3,4 in Figure 3.1), we leave M_{ij} and F_{ij} undefined.

<u>*3.2.2 KMR Framework for Pedigree Data:*</u> We create our robust rare-variant association test for a quantitative trait based on the KMR test of Schifano et al. (2012) and Chen et al. (2013) for association testing of a group of genetic variants with a continuous phenotype allowing for related individuals. As shown by these authors, the KMR test can be implemented in a linear mixed-modeling framework with mean and variance defined through the model:

$$Y_{ij} = \boldsymbol{X}_{ij}^{T}\alpha + h(\boldsymbol{G}_{ij}) + f_i + \epsilon_{ij} \qquad (1)$$

where α is a c×1 vector of coefficients for $\boldsymbol{X}_{ij}$, f_{ij} is the random effect to account for within family correlation, and ϵ_{ij} is the random error term. We further assume that the random effects within a family, f_i, follow a multivariate normal distribution $f_i \sim MVN(0, 2\Phi_i\sigma_{pg}^2)$. Here Φ_i is the kinship matrix for the i^{th} family, elements in Φ_i represent the kinship coefficients between subjects in the i^{th} family, and s^2_{pg}represents the variance due to the shared polygenic effect. We also assume that the random effect e_{ij} is normally distributed with mean 0 and variance s^2_e.

Within equation (1) above, $h(\boldsymbol{G}_{ij})$ is a function of $\boldsymbol{G}_{ij}$ defined through a positive semi-definite kernel function $K(\times,\times)$. It is worth noting that the kernel function, $K(\boldsymbol{G}_{ij}, \boldsymbol{G}_{i'j'})$, measures the genetic similarity between subject *j* in family *i* and subject *j'* in family *i'* and contrasts this similarity to phenotypic similarity between the two subjects. It has been shown that appropriate choice of the kernel can increase the power (Wu, Lee et al. 2011). Frequently used kernels include the identity-by state (IBS) kernel or the linear weighted kernel. The IBS kernel, which takes the form $K(\boldsymbol{G}_i, \boldsymbol{G}_{i'}) = \sum_{j=1}^{s}\left(2 - \left|G_{ij} - G_{i'j}\right|\right)$, measures the genetic similarity as the number of alleles that share by state. It assumes a nonlinear effect of each rare variant and can thus enable the study of epistatic effects. The linear weighted kernel, on the other hand, assumes a linear relationship between the trait and the variants. The kernel takes the form $K(\boldsymbol{G}_i, \boldsymbol{G}_{i'}) = \sum_{j=1}^{s}(w_j G_{ij} G_{i'j'})$. Prior knowledge of the gene can be incorporate by assigning each variant a weight. If prior knowledge is not available, weights can also be calculated as a function of minor allele frequency. Wu et al. (2011) suggests calculating the weights based on a beta distribution, which assigns greater weight to less frequent variants.

It can be easily shown that the estimator of *h* takes the same form as in the linear mixed model with *h* as a random effect (Liu, Lin et al. 2007, Schifano, Epstein et al. 2012):

$$y = Xa + h + f + e. \qquad (2)$$

Thus, the test of whether genotype is associated with the outcome is equivalent to testing whether the random component *h* equals 0 or not. We adopted the variance component score test, which is the locally most powerful test (Lin 1997). For the random effect *h*, it follows an arbitrary distribution with mean 0 and variance $t\boldsymbol{K}$. As a result, the test of whether *h*=0 is equivalent to testing whether $\tau = 0$. The null hypothesis is H$_0$: $\tau = 0$, and the test statistic takes the form:

$$Q = \frac{1}{2}(Y - X\hat{\alpha}_0)\hat{V}_0^{-1} K \hat{V}_0^{-1}(Y - X\hat{\alpha}_0), \tag{3}$$

where all parameters are estimated under the null hypothesis. To obtain the null distribution of Q, we define a projection matrix $P = \hat{V}_0^{-1} - \hat{V}_0^{-1} X(X^T \hat{V}_0^{-1} X)^{-1} X^T \hat{V}_0^{-1}$, such that $PV_0P = P$. Thus, under the null, we have

$$Q = \frac{1}{2} Y^T PKPY = \sum_{i=1}^{N} \lambda_i \chi_{1i}^2, \tag{4}$$

where λ_i are eigenvalues of $\frac{1}{2} D \hat{V}_0^{-1/2} K \hat{V}_0^{-1/2} D$, here $D = I - \hat{V}_0^{-1/2} X(X^T \hat{V}_0^{-1} X)^{-1} X^T \hat{V}_0^{-1/2}$. As χ_{1i}^2 are independently and identically distributed random variables, Q is distributed as an asymptotic mixture of chi-square distributions, and the p-values can be calculated using the Davies method (Davies 1980).

3.2.3 QTDT Framework for General Pedigrees: In the presence of population stratification, association testing of $\boldsymbol{G}_{ij}$ with Y_{ij} in models (1) and (2) may lead to spurious association due to the underlying differences in allele frequencies of the sub-populations. However, for family studies, family members can be used as internal controls, where an expected genotype can be constructed using the family members' information. Tests based on the within-family component (deviation of observed genotype from expected within family) will not be influenced by population structure, even in the most extreme case, where each of the N pedigrees is drawn from a different population. Here, we leverage the work of Abecasis et al. (Abecasis, Cookson et al. 2000) and present the method to calculate transmission scores for individuals in general pedigrees.

The QTDT framework (Abecasis, Cardon et al. 2000) for general pedigrees decomposes a genotype into a between-family component (which is sensitive to population stratification) and a within-family component (which is robust to population stratification). For relative *j* in family i, let $\boldsymbol{B_{ij}}$ and $\boldsymbol{W_{ij}}$ denote vectors of between-family and within-family genotype components for the *s* rare-variant genotypes in $\boldsymbol{G_{ij}}$. Assuming all parents in the pedigree are genotyped, the between-family component for founders (with no ancestors included in the pedigree) will be equal to their observed genotypes, while the between-family component for non-founders at each rare-variant genotype is equal to the average genotype of the between-family components of that individual's parents: such that $\boldsymbol{B_{ij}} = \frac{\boldsymbol{B_{M_{ij}}}+\boldsymbol{B_{F_{ij}}}}{\boldsymbol{2}}$. Using the pedigree in Figure 3.1 as an example, suppose all the individuals in the pedigree are genotyped. Suppressing the family index for ease of presentation, the between-family components for founders 1, 2, 3, 4 are $\boldsymbol{B_1}$=$\boldsymbol{G_1}$, $\boldsymbol{B_2}$=$\boldsymbol{G_2}$, $\boldsymbol{B_3}$=$\boldsymbol{G_3}$, $\boldsymbol{B_4}$=$\boldsymbol{G_4}$ respectfully. For the non-founders in the second generation, the between-family component for individual 5 is $\boldsymbol{B_5} = \frac{\boldsymbol{B_1}+\boldsymbol{B_2}}{\boldsymbol{2}}$, and between-family component for 6 is $\boldsymbol{B_6} = \frac{\boldsymbol{B_3}+\boldsymbol{B_4}}{\boldsymbol{2}}$. For the non-founders in the third generation, the between-family components for individual 7, 8, 9, and 10 are $\frac{\boldsymbol{B}_5+\boldsymbol{B}_6}{2} = \frac{\boldsymbol{B}_1+\boldsymbol{B}_2+\boldsymbol{B}_3+\boldsymbol{B}_4}{4}$. It can be seen that, in the situation where all founders are genotyped, the between-family component of any non-founder is calculated as:

$$\boldsymbol{B}_{ij} = \sum_{f \in F} 2\varphi_{if}\boldsymbol{G}_{if}, \qquad (5)$$

where in the ith family, *f* is the index of founders, $\boldsymbol{G_{if}}$ is the rare-variant genotype vector of the founder, j_{ijf} is the kinship coefficient between individual *j* and founder *f*, and *F* is the set of all the genotyped founders.

In the situation where the parents' genotypes are missing, the between-family component $\boldsymbol{B_{ij}}$ is equal to the average of the genotypes for all sibling of relative *j*. For example in Figure 3.1, if individuals 5 and 6 are not genotyped, then the between-family component for individuals 7, 8, 9, and 10 is $\frac{\boldsymbol{G}_7+\boldsymbol{G}_8+\boldsymbol{G}_9+\boldsymbol{G}_{10}}{4}$. The average of genotypes of siblings in the family is the sufficient statistic for the between-family component (Abecasis, Cardon et al. 2000).

The within-family genotype vector for the *s* rare-variant genotypes $\boldsymbol{W_{ij}}$ is then calculated as the difference between the observed genotype vector and the between-family genotype vector:

$$\boldsymbol{W}_{ij} = \boldsymbol{G}_{ij} - \boldsymbol{B}_{ij} \tag{6}$$

Positive values within $\boldsymbol{W_{ij}}$ indicate excess transmission of the minor (reference) allele, while negative values of $\boldsymbol{W}_{\mathbf{ij}}$ indicate excess transmission of the major allele. As discussed above, the within-family component is not influenced by population substructure; thus, the test on the within-family component is robust to population stratification.

As discussed before, directly testing based on the observed rare-variant genotypes in models (1) and (2) will lead to spurious association in the presence of population stratification. For our robust test, we follow the same approach as in our earlier work (Jiang et al., 2014) and simply calculate $\boldsymbol{W_{ij}}$ as described above, replace $\boldsymbol{G}_{\text{ij}}$ with $\boldsymbol{W}_{ij}$ in equations (1) and (2), and construct our score statistic Q in (3) using $\boldsymbol{W_{ij}}$.

3.2.4 Screening Methods: Although the within-family component has the advantage of robustness to population stratification, constructing tests based only on the within-family genotypic component while ignoring the between-family component reduces power. However, if founders' phenotype and genotype data are available, we can borrow the idea of Purcell et al. (Purcell, Sham et al. 2005) to implement a screening procedure to potentially increase power. Specifically, we use the founders' phenotype and genotype information in the first stage to identify those regions showing strongest signals of association. We can perform such testing using standard burden or variance-component tests for unrelated subjects. We then implement a second stage where we test only the top regions from the first stage using our proposed test in (3) based on the within-family genotypic component; The number of top regions in the second stage can take a value between 1 and the total number of regions. In this project, we assume 10%-50% of the regions enter the second stage. By pre-screening in this manner, we reduce the multiple-testing burden for our robust test thereby increasing power. As the within-family component and the between-family component are orthogonal to each other by design (Abecasis, Cookson et al. 2000), population stratification that can invalidate the first-stage analysis using founders will not invalidate the within-family component test.

3.2.5 Simulation Studies: We evaluate type 1 error rate and power of our method using simulated sequencing data generated by *cosi* (Schaffner, Foo et al. 2005), which has high resemblance with empirical data. To simulate large pedigrees, we first use *cosi* to simulate 5000 haplotypes of European ancestry and 5000 haplotypes of African ancestry. We then randomly draw and pair haplotypes within each population and randomly select one haplotype from each parent to pass down to offspring. Our simulated pedigree has the same structure as Figure 3.1. We assume that there are 10 non-overlapping regions of interest, each 30kb long.

For each family, we simulate phenotype data from a multivariate normal distribution, whose mean and variance vary according to different scenarios. For type I error rate simulations, all 10 regions are null, while for power simulations we randomly select one region of the 10 to harbor causal variation. Rare variants are defined as variants with minor allele frequency (MAF) smaller than 3%. To simulate population substructure, we simulate the outcome for the null model as: $Y_{ij} = gI_{African,\,ij} + f_{ij} + e_{ij}$, where g is the mean trait difference between European and African, and $I_{African,ij}$ is the indicator variable, which is 1 for African individuals and 0 for European individuals. For the power simulations, we let either 5% or 15% of the rare variants in the causal region influence phenotype. For each causal variant, we define the effect size as $b = c´\ |\log_{10} MAF|$, where c is a pre-defined constant. Thus, the outcome is simulated as $Y_{ij} = gI_{African,ij} + b_{ij}´\ G_{ij} + f_{ij} + e_{ij}$.

3.2.6 GAW18 Data: The Genetic Analytic Workshop 18 (GAW18) provides whole genome sequence data for extended pedigrees and phenotypes such as systolic blood pressure (SBP) and diastolic blood pressure (DBP). The dataset was drawn from the T2D-GENES Consortium Project 2; a family-based study that aims to identify low-frequency variants that increase the risk of type-2 diabetes. The original dataset contains whole genome sequences for the odd numbered chromosomes only (chromosomes 1, 3, 5,…,21) for 464 individuals from 20 Mexican American families. The dataset we used in this project contains 959 individuals. 464 of them were directly sequenced by Complete Genomics Inc, while the remaining 495 had sequence data imputed from array-based genotype data by the T2D-GENES Consortium. In addition to SBP and DBP, the dataset also includes information on age, gender, current use of antihypertensive medicine, and current smoking status. We include these phenotypes as covariates in our model. Detailed information about the dataset can be found at Almasy et al. (Almasy, Dyer et al. 2014)

After standard data cleaning procedure removed subjects with missing SBP or DBP measurements, our final dataset contained 855 individuals. Genes were annotated using information from the 1000 Genome Project (http://www.1000genomes.org/). We tested all genes in the 11 odd-numbered chromosomes, where each gene was tested individually. For each gene, we calculated the empirical frequency of the variants within the gene and only performed tests on the rare variants, where a rare variant was defined as having a minor-allele frequency (MAF) less than 3%. We constructed the test statistics using within-family components as defined above.

3.3. Results

3.3.1 Type I Error: We first performed null simulations to show that population stratification can lead to inflated type I error rate for sequencing studies of large pedigrees. Figure 3.2 summarizes the empirical type I error rates of a study with 25 European pedigrees and 75 African pedigrees, each with the same size and family structure as shown in Figure 3.1. We first set the mean trait difference (g) between European and African to be 1 (Figure 3.2 Left) and further increased it to 2 (Figure 3.2 Right). Both figures show that in the presence of population stratification, test statistics constructed on observed genotype have inflated type I error rates (yellow bars in Figure 3.2). As population structure becomes more extreme, the inflation becomes more severe (Figure 3.2 Right). We then performed tests based on our robust test statistics based on our two-stage screening procedure using founders' genotypes and phenotypes. Figure 3.2 shows that testing on the within-family component combined with the screening method leads to appropriate control of the type I error rate in the presence of population stratification.

3.3.2 Power: We next examined power of the proposed robust test. For power simulations, we assume the mean trait different between European and African is 0.25. For each simulation, we randomly drew 25 European pedigrees and 75 African pedigrees from the haplotype pools. We varied the percentage of rare causal variants in the causal region from 5% (Figure 3.3a) to 15% (Figure 3.3b). We also assumed different effect sizes ($b = c' \; |\log_{10} MAF|$) for the causal variants by letting c take the values 0.4, 0.5, and 0.6. Figure 3.3 shows that power increases as the percentage of causal variants in a region increases and as the effect size increases. We next investigated whether the two-stage screening approach using founder information improves power over a within-family analysis that ignores screening. As shown in Figure 3.3, screening on the top 10%-50% of hits can yield noticeable improvements in power over the naïve strategy.

3.3.3 Application to GAW18 Dataset: We used GAW18 data to test for association between DBP/SBP and genes on odd chromosomes. Within each gene, we calculated empirical frequencies of variants and only tested on variants with frequencies smaller than 3%. GAW18 provides longitudinal phenotype information, where SBP and DBP were measured in up to four follow-ups for each subject. We used the baseline measurement to test for association. We also controlled for age, gender, current usage of anti-hypertensive medicine, and current smoking status in our model. The pedigrees are relatively large in the dataset. The median number of individuals in a pedigree is 37 (min 22, max 74). Among the participants, 20.2% of them smoke, 9.4% took medicine, and 57.7% of them are female.

We performed association tests using our robust test. The genome-wide significance level with Bonferroni correction is: $\alpha_{Bonferroni} = 0.05/7034$ =7.1×10^{-6}. We chose the linear weighted kernel and used the Davies method to calculate p-values. Following Wu et al. (2011), the weight is calculated as $w_j \sim Beta(MAF_j, 1,25)$. The results of testing SBP and DBP are summarized in Figure 3.4. As shown in Figure 3.4, we did not observe any genes passing the genome-wide significance level (7.1×10^{-6}, based on Bonferroni adjustment for 7034 genes). At the suggestive level (1×10^{-4}), one gene on chromosome 21is associated with SBP, and one gene on chromosome 7 is associated with DBP. The gene associated with SBP is open reading frame 33 (C21orf33), which is a protein-coding gene and is over-expressed in Down Syndrome

(Yahya-Graison, et al. 2007). LSM5 is associated with DBP at the suggestive level. It has been found that human LSM1 to LSM7 genes were expressed in Hela cells within cytoplasmic foci (Ingelfinger et al., 2002), which contains important factors in the degeneration of mRNA.

3.4. Discussion

In this paper, we presented a framework for rare-variant sequencing studies in large pedigrees. Large pedigrees have several important features that make them ideal for finding traits associated rare variants. Our model, which combines a kernel machine framework for rare-variant analysis with a QTDT framework for general pedigrees, provides a powerful, efficient, and robust way to identify such associations in large pedigree studies. As the test score statistics follows an asymptotically mixed chi-square distribution, the calculation of p-values is much easier compared to other methods. This feature also makes our model applicable to large-scale genetic studies.

We also applied our method on GAW 18 data to identify SBP/DBP associated rare variants. We tested all the genes on odd numbers of chromosomes. This application gives an example that our method can be easily applied to large-scale data. The analysis of a gene takes 70 seconds on a 768 processors running Linux OS with 512 GB or RAM.

The data from GAW18 are based on 20 extended Mexican-American families. For studies that do not have records of participants' geographic origin or studies whose participants are from different origins, our method provides a robust way to perform the test.

In this project, we assumed that rare variants only associated with a single phenotype. However, there is substantial interest in identifying genetic factors with pleiotropic effects that influence multiple distinct phenotypes. Current methods for family data are not well equipped to investigate the effect of pleiotropy. For example, while analyzing GAW18 data, analyses seeking to identify genes simultaneously associated with both SBP and DBP cannot be performed. However Broadaway et al. (2016) provide a framework that can test cross-phenotype effects of rare variants. Their method is based on kernel distance-covariance, whose test statistics also asymptotically follow a mixed chi-square distribution. In contrast to our method presented here, Broadaway et al. focused only on unrelated individuals. In the future, we would like to combine our robust test with the method of Broadaway et al. (2016) to test cross-phenotype effects of rare variants in related individuals.

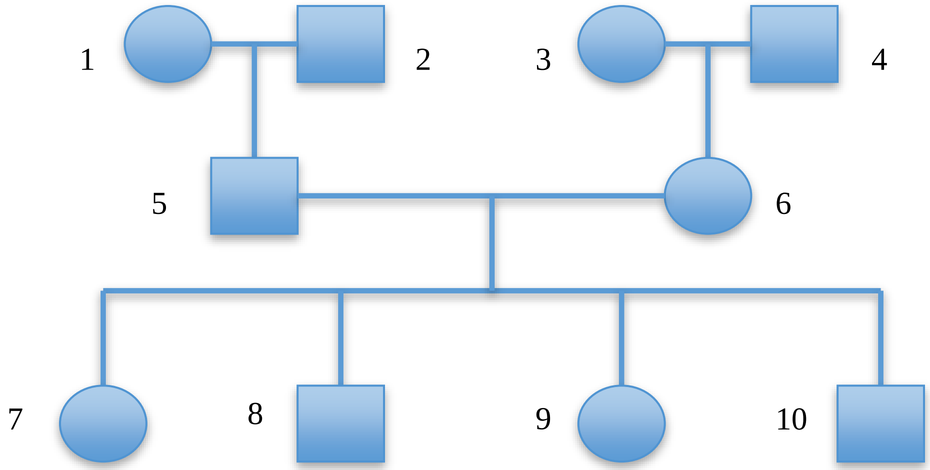


Figure 3.1. Pedigree Structure

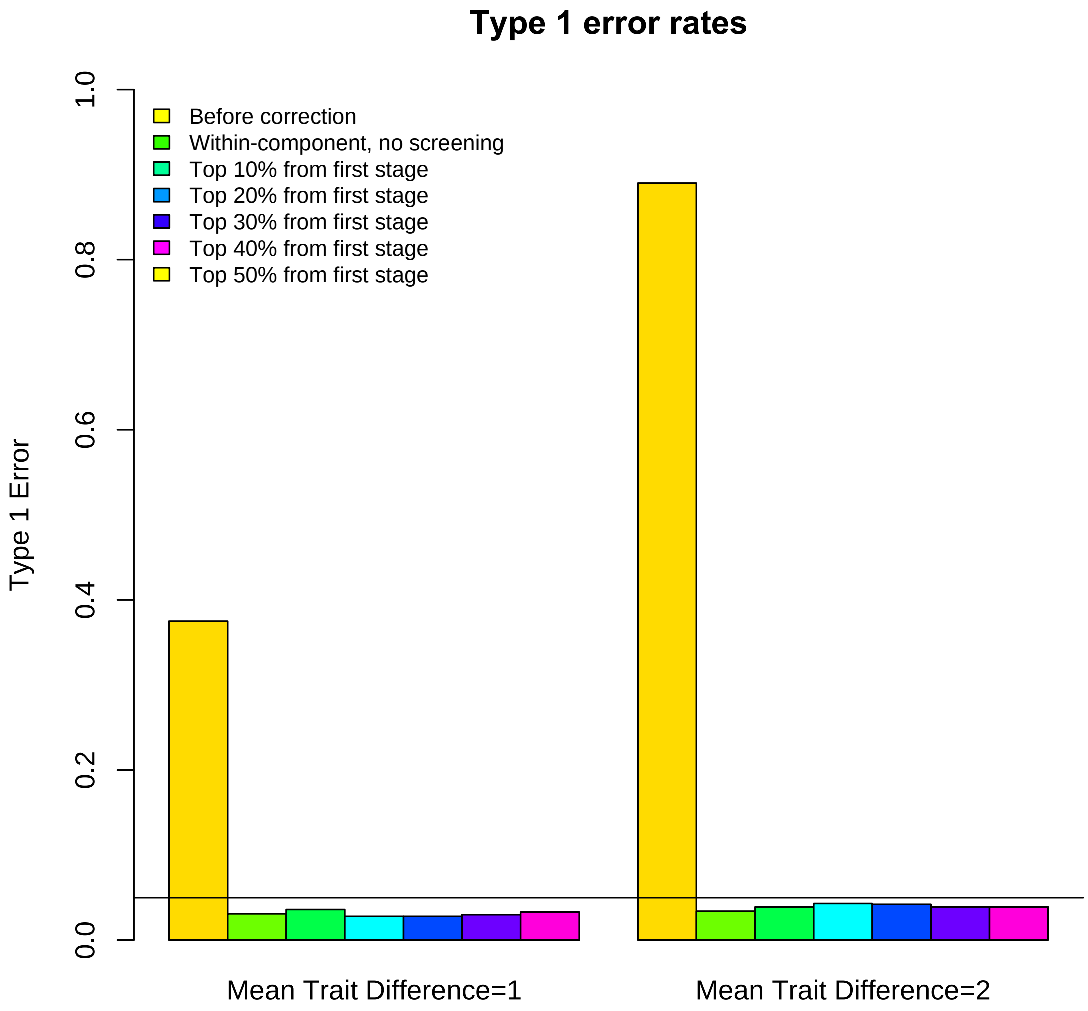
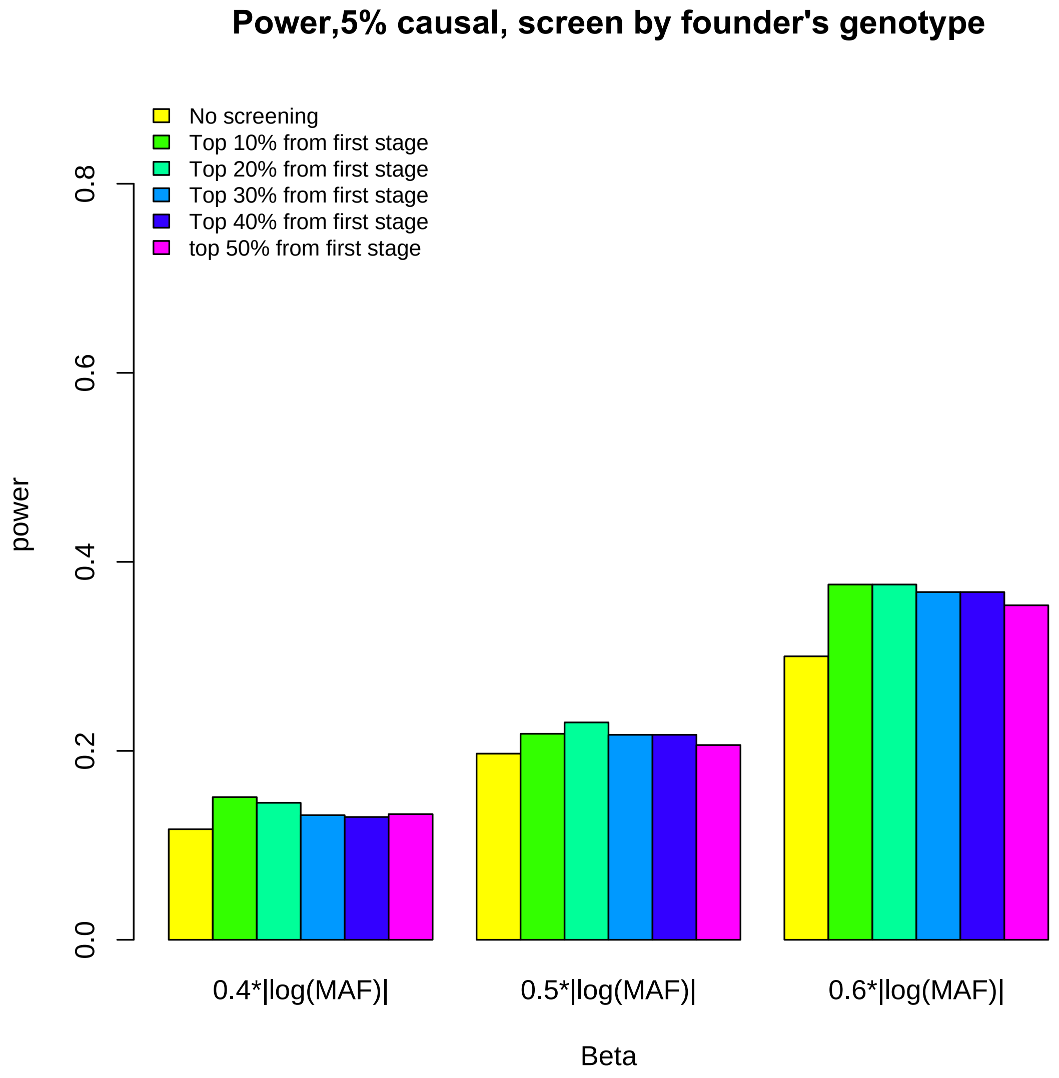


Figure. 3.2. Type 1 Error Rates. Left: Mean trait difference between European and African is 1. Right: Mean trait difference between European and African is 2.
10 30-kb regions are simulated. Yellow bar: Type 1 error rate tested on observed genotype. Others: Type 1 error rate tested on within-family component, with different number of genes at second stage. Black line: y=0.05


Power,5% causal, screen by founder's genotype
No screening
Top 10% from first stage
Top 20% from first stage
Top 30% from first stage
Top 40% from first stage
top 50% from first stage
power
0.0
0.2
0.4
0.6
0.8
0.4*|log(MAF)|
0.5*|log(MAF)|
0.6*|log(MAF)|
Beta

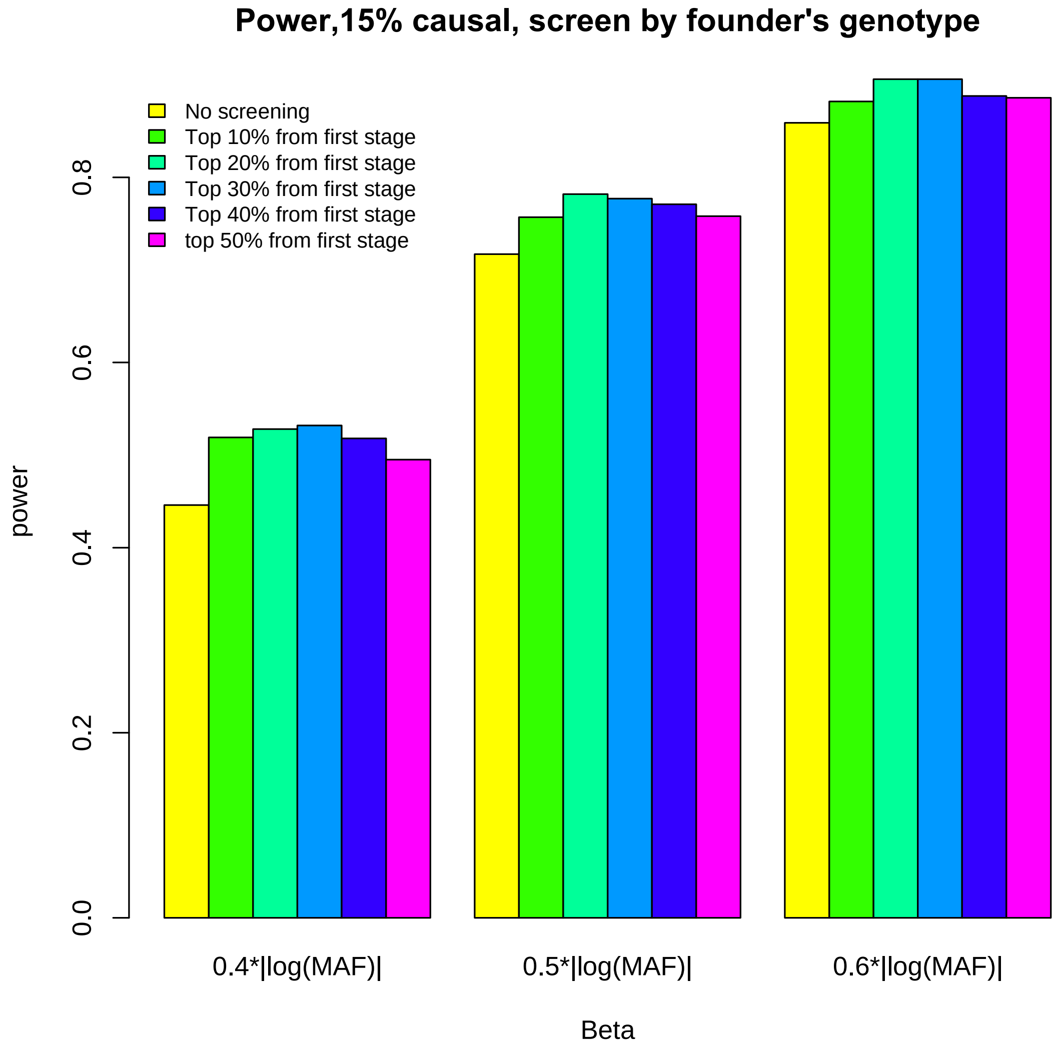


Figure 3.3 Power to detect rare-variant association in large pedigrees. Figure 3a: 5% of rare variants in the causal region are causal variants. Figure 3b: 15% of rare variants in the causal region are causal variants. Yellow bars: Power without screening. Others: Power with screening. Mean trait different between European and African is 0.25. 10-50% regions entered second stage.

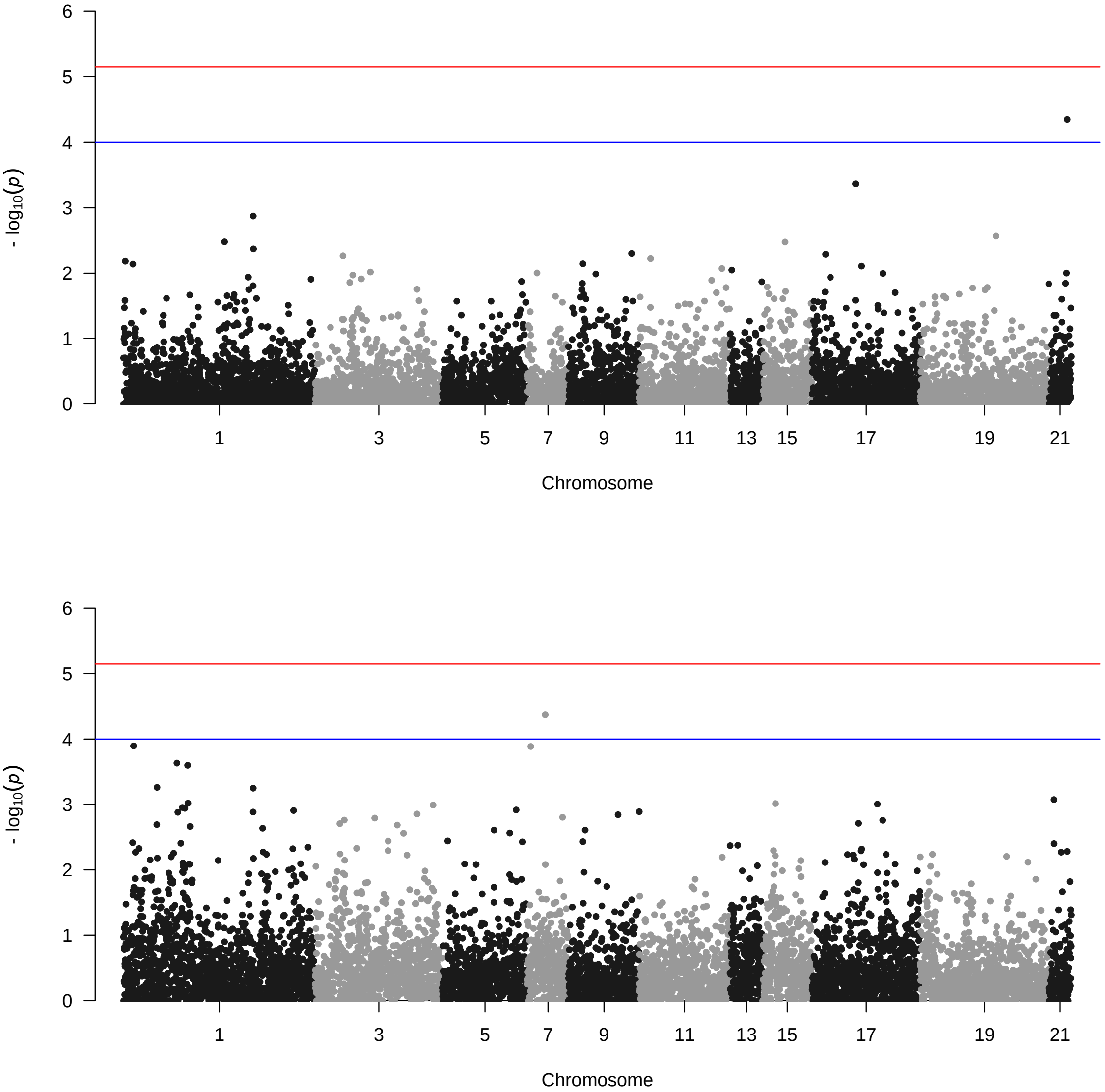


Figure. 3.4. Manhattan plots for GAW18 analyses. Figure 3.4a: Association analyses between SBP and within-family component of genotypes within genes on odd number of chromosomes. Figure 3.4b: Association analyses between DBP and within-family component of within genes on odd number of chromosomes. Red line: Genome-wide significant level ($p<7.1\times10^{-6}$), Blue line: Suggestive significant level ($p<1\times10^{-4}$).

Chapter 4. Powerful and Robust Cross-Phenotype Association Test of Rare Variants in Case-Parent Trios

Abstract

There has been increasing interest in identifying pleiotropic genes within the human genome that influence multiple diverse phenotypes. In the presence of pleiotropy, joint testing of these phenotypes is not only biologically meaningful but also statistically more powerful than univariate analysis of each separate phenotype accounting for multiple testing. While many cross-phenotype association tests exist, the majority of such methods assume samples comprised of unrelated subjects and therefore are not applicable to family-based designs, including the valuable case-parent trio design. In this paper, we describe a robust gene-based association test of multiple phenotypes collected in a case-parent trio study. Our method is based on the kernel distance covariance (KDC) method, where we first construct a similarity matrix for multiple phenotypes and a similarity matrix for genetic variants in a gene; we then test the dependency between the two similarity matrices. The method is applicable to either common variants or rare variants in a gene and resulting tests from the method are by design robust to confounding due to population stratification. We evaluated our method through simulation studies and observed that the method is substantially more power than standard univariate testing of each separate phenotype. We also applied our method to phenotypic and genotypic data collected in case-parent trios as part of The Genetics of Kidneys in Diabetes (GoKinD) study and identified a genome-wide significant gene demonstrating cross-phenotype effects that was not identified using standard univariate approaches.

4.1 Introduction

Pleiotropy has been an increasingly important topic in the genetic association literature since power may be gained by studying the joint influence of a single gene on multiple phenotypes (Solovieff, Cotsapas et al. 2013, Kocarnik and Fullerton 2014). A pleiotropic effect occurs when a single molecular function affects multiple biological processes (He and Zhang 2006). A large number of genes demonstrate pleiotropic effects: a 2011 review article examined NHGRI (National Human Genome Research Institute)'s catalog of common variants and found 233 (16.9%) genes that were significantly associated with multiple traits (Sivakumaran, Agakov et al. 2011). The importance of studying genes or genetic variants with cross-phenotype effects also extends to secondary phenotypes. For example, diabetes studies typically measure Systolic blood pressure (SBP), diastolic blood pressure (DBP), high-density lipoprotein (HDL), body mass index (BMI) as secondary phenotypes. Joint analysis of these phenotypes not only provides biological insight but also increases effective sample size and subsequently improves power (Diggle 2002, Galesloot, Van Steen et al. 2014).

While a variety of statistical methods exist for testing the association between genetic variants and multiple phenotypes (Zhao and Thalamuthu 2011, Maity, Sullivan et al. 2012, O'Reilly, Hoggart et al. 2012, Yang and Wang 2012, Guo, Liu et al. 2013, Schifano, Li et al. 2013, Galesloot, Van Steen et al. 2014, Broadaway, Cutler et al. 2016), several limitations remain in this area. The majority of existing methods test association between individual genetic variants and multiple phenotypes. However, gene-based tests that jointly consider multiple variants in a region of interest have several advantages over single marker tests. First, gene-based tests combine signals from variants within a gene together, which makes them particularly appealing for rare-variant sequencing studies where individual rare variants may be difficult to detect. Second, gene-based tests can adopt dimension-reduction tools, such as kernel machines, to lower the multiple-comparison burden and accommodate complex relations between markers as well as non-linear effect between genetic variants and phenotypes. Several approaches have used gene-based testing to improve upon traditional single-marker tests but typically focused on a single phenotype rather than multiple phenotypes.

To address this issue, Maity et al. (Maity, Sullivan et al. 2012) proposed a gene-based test of multiple phenotypes for common variants using multivariate kernel machine regression (MV-KMR). As a multivariate version of kernel machine regression, the estimation of parameters in the model has the same form as multivariate linear mixed model. Similar to other KMR methods (Epstein and Kwee 2007, Kwee, Liu et al. 2008, Wu, Kraft et al. 2010, Wu, Lee et al. 2011), they also adopted a variance-component score test to reduce multiple testing burden. However, limitations arise as the MV-KMR method 1) is only applicable to continuous traits 2) only allows linear correlations between phenotypes 3) requires computational intensive permutation procedures to calculate p-values. To tackle this issue, Broadaway et al. (Broadaway, Cutler et al. 2016) proposed a method for gene-based testing of multiple rare variants using a kernel distance covariance (KDC) framework. Unlike the method of Maity et al., this method not only defines a similarity matrix among genetic variants in a gene but also defines a similarity matrix for phenotypes. The KDC framework then tests whether the individual elements of the phenotype similarity matrix are independent of the individual elements of the genotype similarity matrix. This method, called GAMuT, can accommodate both binary and continuous traits, and the test statistics asymptotically follows a mixture of chi-square distributions, which makes the calculation of p-value straightforward.

While the methods of Maity et al. (2012) and Broadaway et al. (2016) are valuable, they are also limited to studies of unrelated subjects. It would be useful to create a gene-based test of multiple phenotypes applicable to family designs, particularly the valuable case-parent trio design. An attractive feature of case-parent trio designs is that, by leveraging the QTDT framework (Abecasis, Cardon et al. 2000), one can create tests that by design are robust to population stratification. Population stratification occurs when samples originate from multiple populations with different disease prevalence and different distributions of minor allele frequencies. It can lead to power loss and substantially inflated type I error rates when left unaddressed (Epstein, Duncan et al. 2012, Jiang, Epstein et al. 2013).

In this manuscript, we propose a novel gene-based test for cross-phenotype association testing in case-parent trio studies that is robust to population stratification. We base our approach on the kernel distance-covariance (KDC) framework utilized in the GAMuT test (Broadaway et al. 2016) but replace the observed genotype information in that test with robust within-family genotypic information derived from the QTDT framework (Abecasis et al. 2000). In the following sections, we will first introduce how we construct our test statistics using the KDC framework and how we make these statistics robust to population stratification through incorporation of a QTDT framework. We evaluate our method using simulations and further compare results of our method with gene-based testing of univariate phenotypes. We will also apply our method to a real GWAS case-parent trio study of type 1 diabetes-related phenotypes collected by The Genetics of Kidneys in Diabetes (GoKinD) study. Finally, we will conclude with a summary of our findings and discussion of future extensions.

4.2 Materials and Methods

4.2.1 Notation

We assume a sample of N case-parent trios (parents and offspring) that are genotyped in a gene or region of interest and are measured for multiple phenotypes. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2} \ldots Y_{iL})$ denote the L phenotypes for the offspring in family i, where i=1,2,…N. Let $\mathbf{G}_i = (G_{i1}, G_{i2,\ldots} G_{iS})$, an S×1 vector, denote the S genotyped variants for the same offspring, where G_{is} is coded as number of minor alleles that the offspring possesses at site s. The variants in the gene can consist of either common variants (Kwee et al. 2008) or rare variants (Wu et al. 2011) We further define $\mathbf{X}_i = (X_{i1}, X_{i2,\ldots} X_{iC})$ as a C×1 vector of covariates for the offspring. Let $\mathbf{Y=(Y_1, Y_2,\ldots,Y_N)^T}$ denote the N x L matrix of offspring phenotypes in the dataset and let $\mathbf{G=(G_1, G_2,\ldots,G_N)^T}$ denote the N x S matrix of offspring genotypes.

4.2.2 Kernel Distance Covariance Test of Independence

We wish to create a robust association test between phenotypes **Y** and gene-based genotypes **G** for case-parent trios using the Kernel Distance Covariance (KDC) framework (Gretton, Fukumizu et al. 2007). To do this, we leverage the work of Broadaway et al. (2016), who showed how to create such a KDC-based test (Gene Association with Multiple Traits, named "GAMuT") for gene-based testing of multiple phenotypes in population-based studies. We derive their approach first and then discuss how we leverage their work to develop the robust test of cross-phenotype effect for case-parent trio studies

The GAMuT test of Broadaway et al. (2016) is based on the independence test between kernels on reproducing kernel Hilbert spaces (RHKS) first introduced by Bach et al. (Bach and Jordan 2002). For Hilbert spaces, Bach et al. showed that the canonical correlation of two kernels equals zero if and only if the two variables are independent. Based on this finding, Gretton et al.

(Gretton, Fukumizu et al. 2007) extended the test to use the Hilbert-Schmidt norm as a measure to test the independence between two kernels. The advantages of their method are: 1) the calculation is straightforward and computational complexity is proportional to the square of sample size, which is appealing for high-dimensional genomics data, and 2) the test statistics asymptotically follow a mixture of chi-square distributions, which makes the calculation of p-value efficient to derive. These characteristics make KDC ideal for testing independence between a kernel similarity matrix based on multiple phenotypes, **Y** and a kernel similarity matrix based on multiple variants in **G**, the gene/region of interest.

Based on these findings, Broadaway et al. (2016) first constructed a similarity matrix between phenotypes (**Y**) and a similarity matrix between genotypes (**G**), and then tested for dependency between the two similarity matrices. Let **P** denote the phenotype similarity matrix, where commonly used similarity methods to construct **P** include the construction of a projection matrix ($\mathbf{P}=\mathbf{Y}(\mathbf{Y}^{T}\mathbf{Y})^{-1}\mathbf{Y}^{T}$ (Wessel and Schork 2006)) or linear kernels ($\mathbf{P}=(\mathbf{Y}_{ij},\mathbf{Y}_{i'j'})=\sum_{l=1}^{L} Y_{ijl} Y_{i'jl}$). Let **K** denote the genotype similarity matrix, where commonly used methods to construct **K** include construction of an IBS kernel ($\mathrm{K}(G_i, G_j) = \sum_{s=1}^{S} IBS(G_{is}, G_{js})/2S$, $IBS(G_{is}, G_{js})$ calculates the number of alleles shared identical by descent for subject *i* and *j* at the *s*th variant) or weighted linear kernel ($\mathbf{K}=\mathbf{GTG}^{T}$, where $\mathbf{T}=\mathrm{diag}(weight_1, weight_2, \ldots weight_s)^{T}$ is a diagonal matrix with relative weight for each variant on the diagonal) (Wu, Kraft et al. 2010). Choice of kernels depends on prior assumptions of the relationships between phenotypes or genotypes; appropriate choice of the kernel can increase power (Kwee, Liu et al. 2008, Schaid 2010, Wu, Kraft et al. 2010).

The GAMuT test of Broadaway et al. (2016) is based on the test of Bach et al (Bach and Jordan 2002) which relies on centered kernels for inference. Therefore we further define a centering matrix $\mathbf{H} = (\mathbf{I} - 1_N 1_N^T / N)$ (Schölkopf, Smola et al. 1998) , where **I** is a *N* dimensional identity matrix and 1_N is a N×1 vector of 1, such that **K**c=HKH and **P**c=HPH are centered matrices. Following the above notation, GAMuT is constructed as

$$Q = \frac{1}{N} trace(\mathbf{KHPH}). \qquad (1)$$

Under the null hypothesis of no association, the test statistic follows the asymptotic mixed chi-square distribution defined as $\frac{1}{N^2} \sum_{m,n} \lambda_{P_c,m}\, \lambda_{K_c,n} z_{mn}^2$, where $\lambda_{K_c,n}$ is the *n*th non-zero eigenvalue of **K**c, $\lambda_{P_c,m}$is the *m*th non-zero eigenvalue of **P**c , and z_{mn} are independent and identical distributed standard normal variables. GAMuT uses Davies' method (Davies 1980) to analytically calculate the p-value of the GAMuT, thereby avoiding the need for computationally expensive permutations for inference.

4.2.3 KDC Test for Case-Parent Trios

As we discussed in the introduction, the limitation of the original GAMuT test based on KDC is it cannot be used to construct robust cross-phenotype association tests for use in case-parent trio studies. To address this issue, we propose a robust modification of the GAMuT test statistic for trio data by integrating the QTDT (Abecasis, Cardon et al. 2000) framework within the KDC framework. The QTDT framework decomposes the observed genotype of a trio offspring $\boldsymbol{G}_{is}$ into a between-family component, $\mathbf{B}_{is}$, and a within-family component, $\mathbf{W}_{is}$. $\mathbf{B}_{is}$ is calculated as the average genotype of the offspring's parents, which can be viewed as the mean genotype of the founder's sub-population and is thus sensitive to population stratification. The within-family component $\mathbf{W}_{is}$ is calculated as $\mathbf{G}_{is}$- $\mathbf{B}_{is}$. Because $\mathbf{W}_{is}$ can be viewed as the deviate from the sub-population mean, it is thus robust to population stratification. Thus, to create a robust gene-based test of multiple phenotypes for use in case-parent trios, we first use parental genotypes to calculate $\mathbf{B}_{is}$ and $\mathbf{W}_{is}$ for each variant in the gene and then replace **G** with **W** in the construction of **K**. By doing this, our resulting test is robust to population stratification.

4.2.4 Simulations

We first evaluate the type I error rate and power of our method using simulated data. For each simulation, we use the coalescent simulator *cosi* (Schaffner, Foo et al. 2005) to simulate a pool of 5000 European haplotypes and 5000 African haplotypes, each 30kb long. *Cosi* uses a coalescent model to simulate haplotypes based on empirical patterns of genetic variation observed in different ancestral populations. To simulate sequencing data for trios, we first randomly select haplotypes within each population and pair them for the father and mother of the trio. We then randomly select one haplotype from each parent to form the offspring's haplotypes. In order to examine whether our method is robust to population stratification, we assume the sample consists of 75% trios of African origin and 25% trios of European origin. We further assume the mean trait difference between European and African subjects is 0.3 (R^2: 0.69). We consider tests both of common variation as well as rare variation in a gene. Rare variants are defined as variants with minor allele frequency less than 3%. Common variants are defined as variants with minor allele frequency greater than 5%.

For type I error rate simulation, we assume 6 phenotypes are recorded, and the residual correlation among them follows a multivariate normal distribution with pairwise-trait correlation sampled from a uniform distribution (0,0.3). To examine the robustness of our model to confounding due to population stratification, we assume 2, 4, or 6 phenotypes are affected by population stratification. We constructed the test both using the observed genotype information (corresponding to the GAMuT test) and just using the robust within-family component.

For power simulations, we also assume 6 phenotypes are recorded. As discussed in Kocarnik et al. (Kocarnik and Fullerton 2014), pleiotropy involves both highly correlated phenotypes and phenotypes that are very diverse. To simulate this, we again assume that phenotypes follow a multivariate normal distribution but that the pairwise trait correlations are drawn from a uniform distribution. We separately consider scenarios reflecting low correlation (pairwise correlation of phenotypes drawn from a uniform distribution with bounds (0,0.3), medium correlation (0.3,0.5), or high correlation (0.5,0.7). We also varied the number of phenotypes that are truly associated: we assume either 2 or 4 of the 6 phenotypes are associated with the causal variants in the region. For the rare variants test, we assume that 5% of rare variants in the region are causal. For each causal variant, we simulated the effect as $\beta = (0.4 + N(0,0.1)) \times |log_{10}^{MAF}|$ such that less frequent alleles have larger effects on the outcome (Wu, Kraft et al. 2010). For the common variants test, we assume that 5% of common variants in the region are causal. For each causal common variant, we assume a fixed effect where $\beta = log_e^{1.5}$. We compare our method with RF-KMR, a robust univariate gene-based test of a continuous phenotype using within-family information (Jiang, Epstein et al. 2013) that does not consider multiple phenotypes simultaneously. As RF-KMR tests each phenotype individually, it is necessary to adjust for the 6 tests performed. As the phenotypes are correlated, direct application of the Bonferroni correction will be conservative. We instead calculate the number of effective tests, $L_{effective}$, as the number of principal components able to explain 90% of the variance of phenotypes. We then calculate the adjusted threshold as $\alpha_{effective} = \alpha / L_{effective}$. Compared to traditional Bonferroni correction, this threshold will achieve appropriate Type I error rate and increased power.

4.2.5 GoKinD Data Analysis

As a real-data application example, we applied our method to common variants from a case-parent trio GWAS study of type 1 diabetes from the GoKinD study(Mueller, Rogus et al. 2006, Pezzolesi, Poznik et al. 2009). While GoKinD was initially designed to identify genes associated with diabetic nephropathy in type 1-diabetes patients, the study collected additional phenotypes that can potentially provide more insights in this line of research, such as systolic blood pressure (SBP), diastolic blood pressure (DBP), high-density lipoprotein (HDL), and body mass index (BMI). The study made available phenotype and genotype data on 584 parent-offspring trios on dbGaP (phs000018.v2.p1 and phs000088.v1.p1). All subjects were genotyped using the Affymetrix Mapping 500K array. We used the annotation file from the 1000 Genomes Project to identify common SNPs that fell within known genes. After excluding genes with less than two common variants, 9,647 genes and 131,366 SNPs were included in our analysis. We used our novel cross-phenotype test to test the association between the 9647 genes and SBP, DBP, HDL, BMI. We also adjusted for important covariates in our model: gender, age, renal function status (proteinuric, dialysis, renal transplant or other), smoking status, insulin intake (yes or no), anti-hypertension drug intake (yes or no), lipid lowering medication intake (yes or no). We applied both our method and univariate RF-KMR testing to the dataset.

4.3 Results

4.3.1 Type I error rate

We first applied our method to 10,000 simulated datasets that were subjected to confounding. For each simulation, we sampled 500 trios (125 European and 375 African) from the pool of 10,000 simulated haplotypes. We constructed the test statistics using both the observed genotype (sensitive to population stratification) and the (robust) within-family component. We chose the weighted linear kernel to form the genotype similarity matrix, where weight was generated through the beta distribution density function evaluated at the minor allele frequency: weight~beta(MAF, 1, 25) (Wu, Lee et al. 2011). We further chose the projection matrix to form the phenotype similarity matrix.

We summarized type I error rates using Quantile-Quantile (q-q) plots in Figure 4.1 (rare variants) and Figure 4.2 (common variants). In the presence of population stratification, the distribution of p-values for the GAMuT test of observed genotypes significantly deviated from the expected uniform distribution (top panels of Figures 4.1 and 4.2). As more phenotypes are affected by stratification, the deviation becomes more extreme. However, our method (bottom panels of Figure 4.1 and 4.2) yields the expected distribution of p-values under the null under all circumstances, suggesting that the correct type I error rate is achieved at all significance levels.

4.3.2 Power

Our type I error rate simulations showed that our method using the within-family component is robust to population stratification. In this section, we will evaluate the power of our method and compare the results with the robust univariate test, RF-KMR (which is the within-family KMR method of Jiang et al. (2014)). We applied our method to 5,000 simulated datasets. Similar to above, each simulation sampled 500 trios (125 European and 375 African) from the pool of haplotypes.

As described in Methods, we simulated such that 5% of rare variants in the region were causal and we varied the number of phenotypes associated with the causal variants. We use the project matrix as the phenotypic similarity matrix and weighted linear kernel as genotypic similarity matrix, where weight~Beta(MAF,1,25). Simulation results are summarized in Figure

4.3. Under all circumstances, our method (orange bars) outperforms univariate kernel machine regression (green bars). Our method can capture the correlation between phenotypes: power increases as correlation between phenotypes increases (Figure 4.3, left to right), while the univariate test cannot exploit this information. Power also increases as the number of phenotypes associated with the outcome increases for weakly correlated phenotypes (Figure 4.3, left to right). For the common variants test, we also simulated such that 5% of common variants in the region were causal, but with an effect size (log_{10}^{MAF}) smaller than that assumed for rare variants. We use the weighted linear kernel as genotypic similarity matrix, where the weight is calculated as $\frac{1}{\sqrt{MAF}}$ as suggested in Wu et al. (Wu, Kraft et al. 2010). For rare variant tests, we also observed that our method achieves higher power compared to the univariate test under all simulation settings (Figure 4.4).

4.3.3 GoKinD Data Analysis

Using genotype and phenotype data from the GoKinD study available from dbGaP (see Web Resources and Acknowledgements), we tested the phenotypes SBP, DBP, HDL, and BMI for association with common variants. We removed variants whose missing rate is larger than 5%. For each phenotype, we replaced missing values with the median value of the phenotype. The final sample consisted of 544 trios with genotypes on 131,366 common variants from 9647 genes, with a median of 13.6 variants in each gene. We analyzed the data using our method, which tests all four phenotypes simultaneously, and the univariate RF-KMR method, which tests each phenotype individually. For both tests, we tried using both linear kernels and weighted linear kernels to form the genetic similarity matrix. We also adjusted for age, gender, renal function status, smoking status, insulin intake, anti-hypertension drug intake, lipid lowering medication intake. To adjust for covariates in our method, we first regress each phenotype separately on the covariates, and then use the residuals from each regression to form the phenotype similarity matrix (linear kernel). We assume a Bonferroni-adjusted genome-wide significance level of 0.05/9647≈5×10^{-6}, and a suggestive level of 5×10^{-4}. As the RF-KMR method tests each phenotype individually, we adjusted for multiple testing through the following procedure: we first find the minimum p-value among the 4 tests, and then multiply it by an estimate of the effective number of tests (the number of principal components that can explain 90% variation of the 4 phenotypes). This procedure is less conservative than Bonferroni correction, allowing for a fairer comparison between methods. We construct our test statistics using both the observed genotype and the robust within-family component. A QQ plot from the observed genotype test showed that there is no inflation in the test statistics (Supplementary Figure 4.1.). This is unsurprising, as over 96% of samples are white.

We first formed the genetic similarity matrix using linear kernel and summarized the results in Manhattan plots and QQ plots (Figures 4.5 and 4.6). By utilizing information from the correlation between phenotypes, our method systematically yields smaller p-values than RF-KMR. Using our method, we identified a gene Vacuolar Protein Sorting 41 (*VPS41,* containing 47 SNPs in our data) on chromosome 7 that passes the genome-wide significance threshold (Supplementary Figures 4.2 and 4.3). We also formed the genetic similarity matrix using a weighted linear kernel, where weight is calculated as $-log_{10}^{MAF}$and, for these weighted analyses, *VPS41* nearly approached genome-wide significance.

4.4 Discussion

In this paper, we introduced a method of identifying pleiotropic genes in related individuals. Our method is a gene-based method that can incorporate prior information about the gene and suits for testing both rare and common variants. The test is a non-parametric test based on KDC framework. The framework has several appealing features that make it ideal for high-dimensional data as 1) the calculation of test statistic in very intuitive, 2) the test statistic follows an asymptotic distribution which makes the calculation of p-value very easy. Our method further incorporated QTDT framework with the KDC framework to make the model robust to population stratification and applicable to related individuals. We performed simulation studies on both the rare variants and the common variants; both studies showed that our method is robust to population stratification and processes more power comparing to the univariate test we previously developed. We will make the code available through the web resource (http://genetics.emory.edu/labs/epstein/software/).

Applying our method to publicly available data from the GoKinD Study, we identified a gene not previously reported to associate with diabetes-related phenotypes (DBP, SBP, HDL, and BMI). *VPS41* is a member of Vesicle medicated protein sorting family, which plays an important role in segregation of intracellular molecules into distinct organelles. Previous work has shown that VPS41 associates with class C VPS proteins to form the complete homotypic fusion and protein sorting (HOPS) compelx (Plemel, Lobingier et al. 2011). Expression studies have shown that VPS41 is potentially involved in the formation and fusion of transport vesicles from the Golgi.

Our method tests on either rare or common variants in a gene. There is a growing interest in examining the combined effect of rare and common variants. Ionita-Laza et al. (Ionita-Laza, Lee et al. 2013) proposed such framework for the uni-variate test. In their paper, they constructed the combined test as the weighted sum of the test statistics from rare and common variants test, where the weight can be assigned using prior knowledge. It should be easy to incorporate their method into our framework to test the combined effect of rare and common variants on multiple phenotypes.

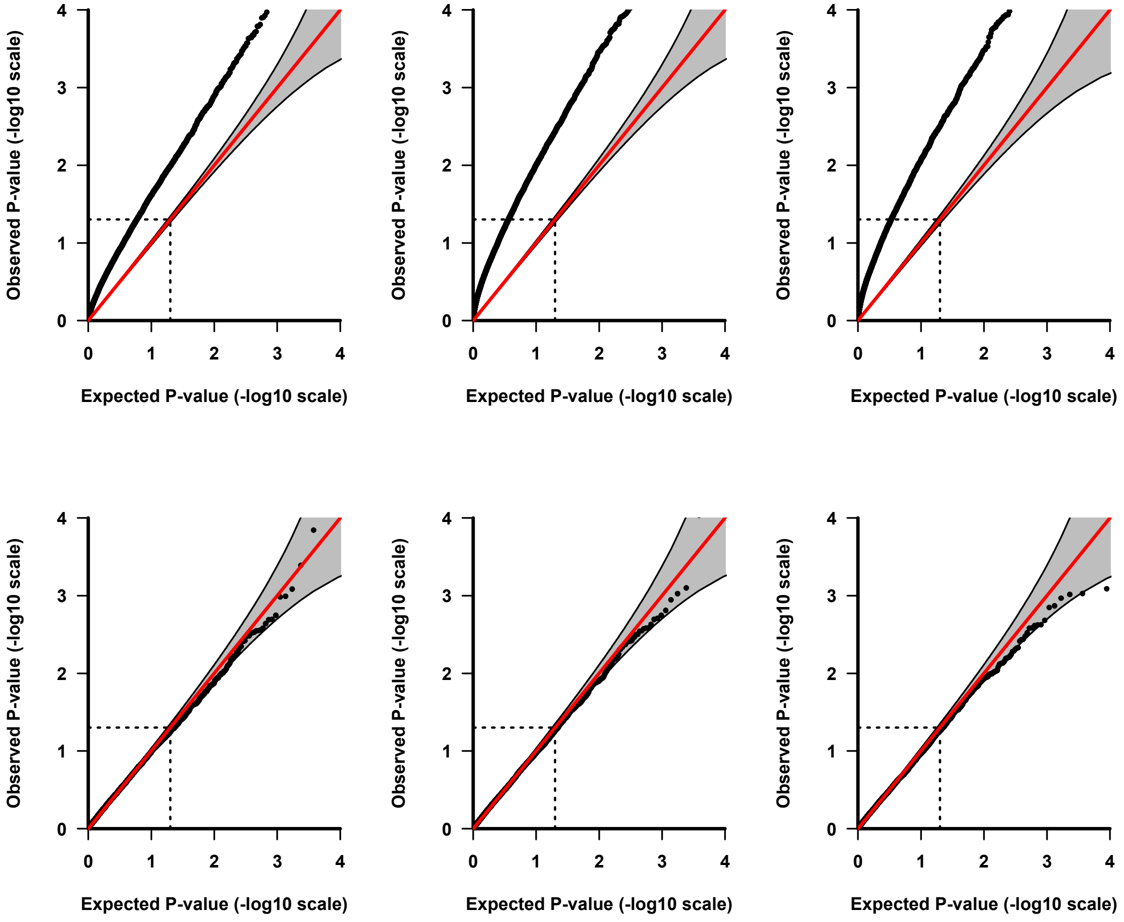


Figure 4.1 Q-Q plots of p-values for gene-based tests of rare variants with 6 null phenotypes using 10,000 simulations. Top panel: Tests on observed genotype. Bottom panel: Tests on within-family component. Left panel: 2 phenotypes affected by population stratification. Middle panel: 4 phenotypes affected by population stratification. Right panel: 6 phenotypes affected by population stratification

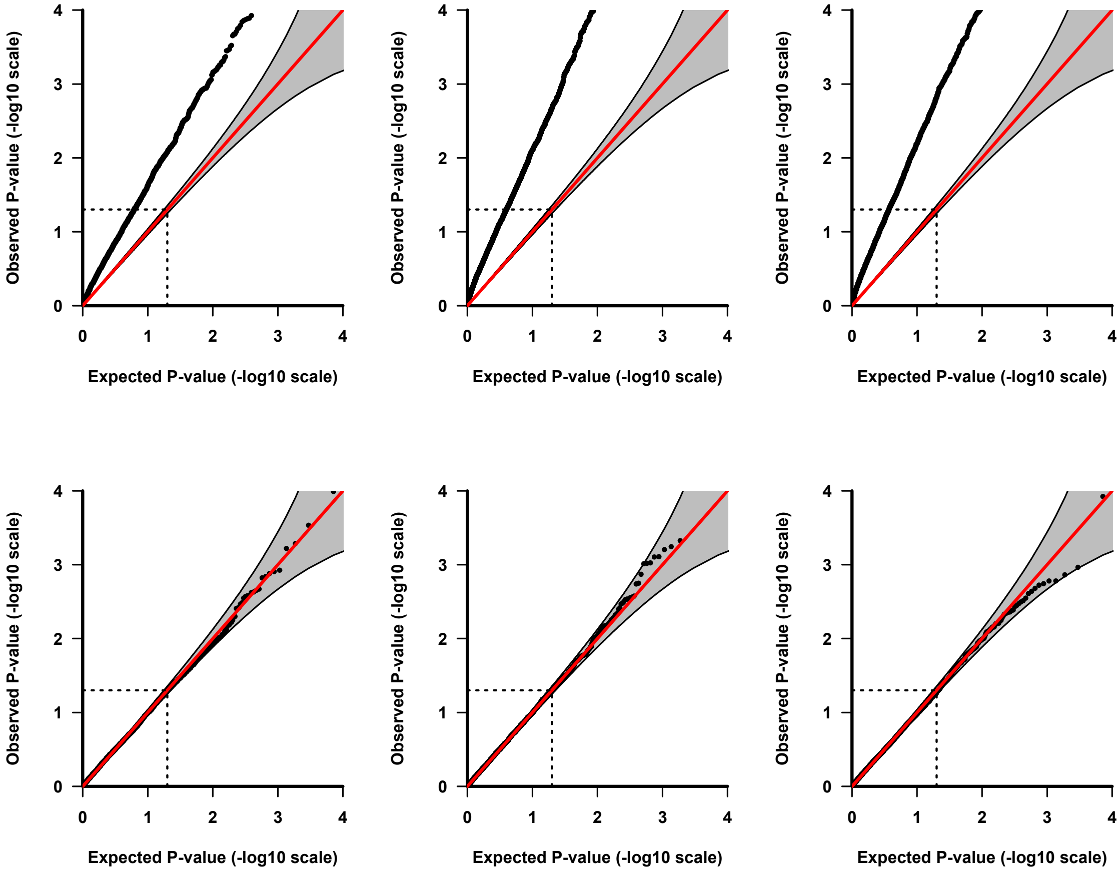


Figure 4.2 Q-Q plots of p-values for gene-based testing of common variants with 6 null phenotypes using 10,000 simulations. Top panel: Tests on observed genotype. Bottom panel: Tests on within-family component. Left panel: 2 phenotypes affected by population stratification. Middle panel: 4 phenotypes affected by population stratification. Right panel: 6 phenotypes affected by population stratification

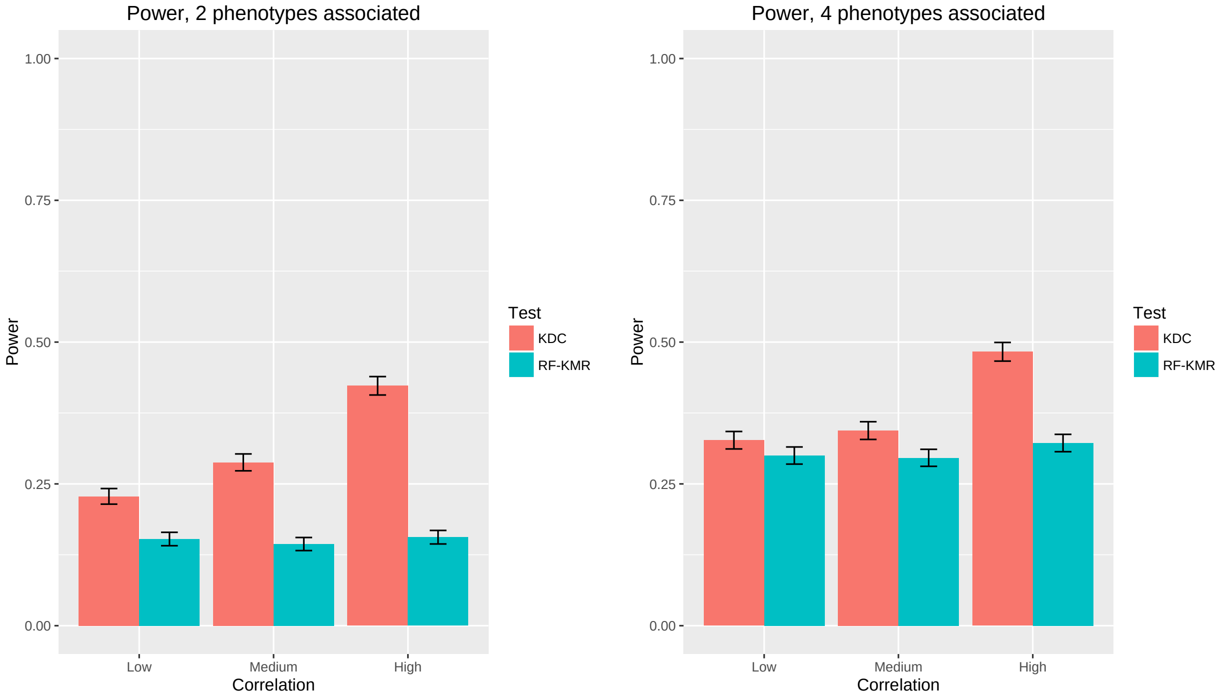


Figure 4.3 Power for gene-based testing of rare variants with 6 phenotypes. Orange bar: Power of cross-phenotype test using KDC. Green bar: Power of test using univariate RF-KMR with adjusted Bonferroni to correct for multiple comparisons. Left panel, 2 phenotypes associated with the causal rare variants. Right panel, 4 phenotypes associated with the causal rare variants. All tests are constructed using robust within-family component. The results are based on 5000 simulations

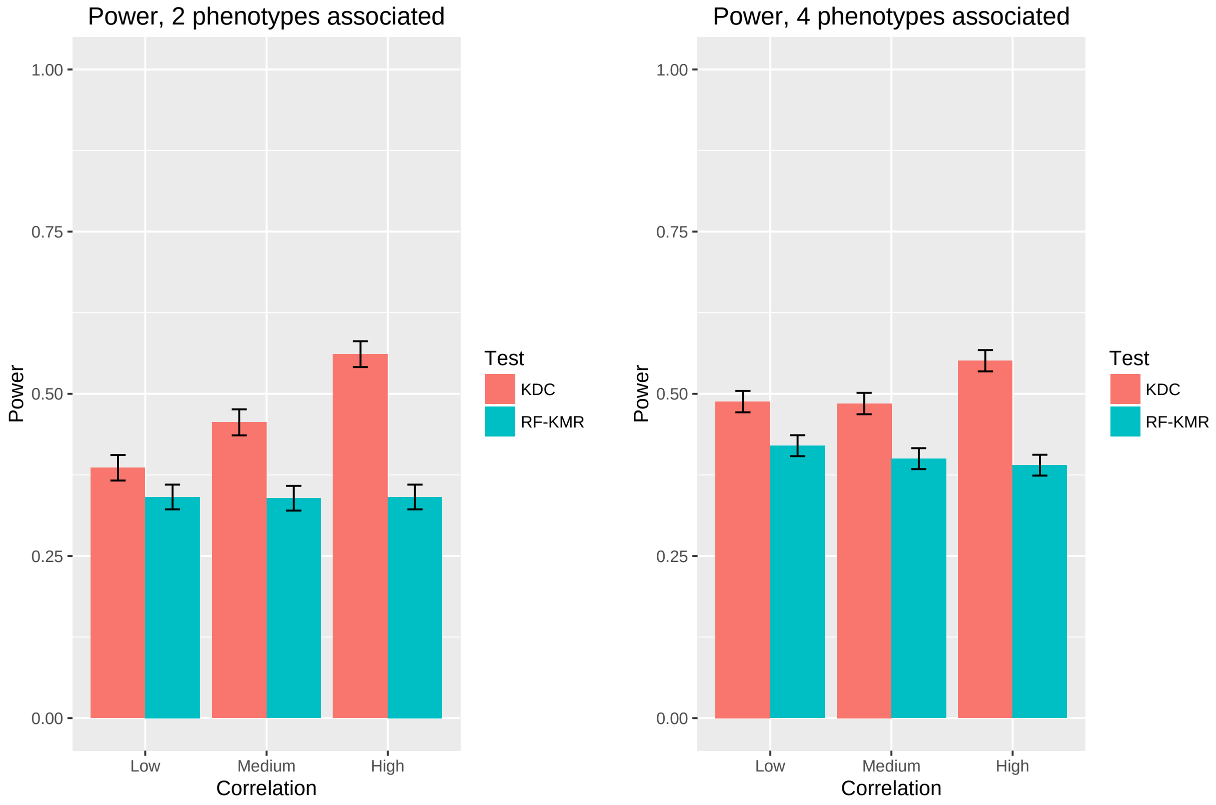


Figure 4.4 Power for gene-based testing of common variants with 6 phenotypes. Orange bar: Power of cross-phenotype test using KDC. Green bar: Power of test using univariate RF-KMR with adjusted Bonferroni to correct for multiple comparisons. Left panel, 2 phenotypes associated with the causal rare variants. Right panel, 4 phenotypes associated with the causal rare variants. All tests are constructed using robust within-family component. The results are based on 5000 simulations

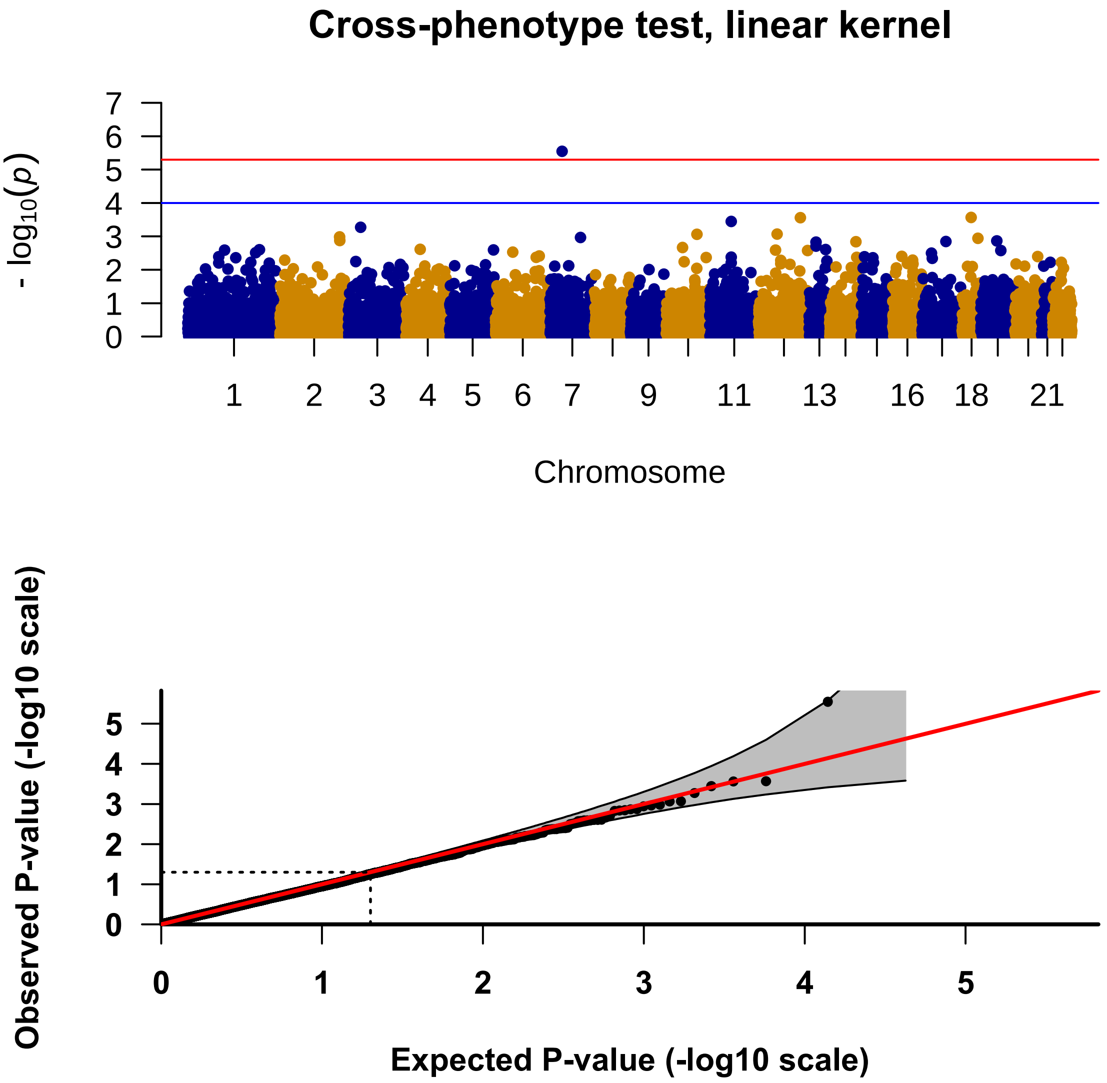


Figure 4.5 Cross-phenotype test on GoKinD data. Top: Manhattan plot for cross-phenotype test with linear kernel. Red line: genome-wide significance level. Blue line: suggestive level. Bottom: Quantile-Quantile Plot

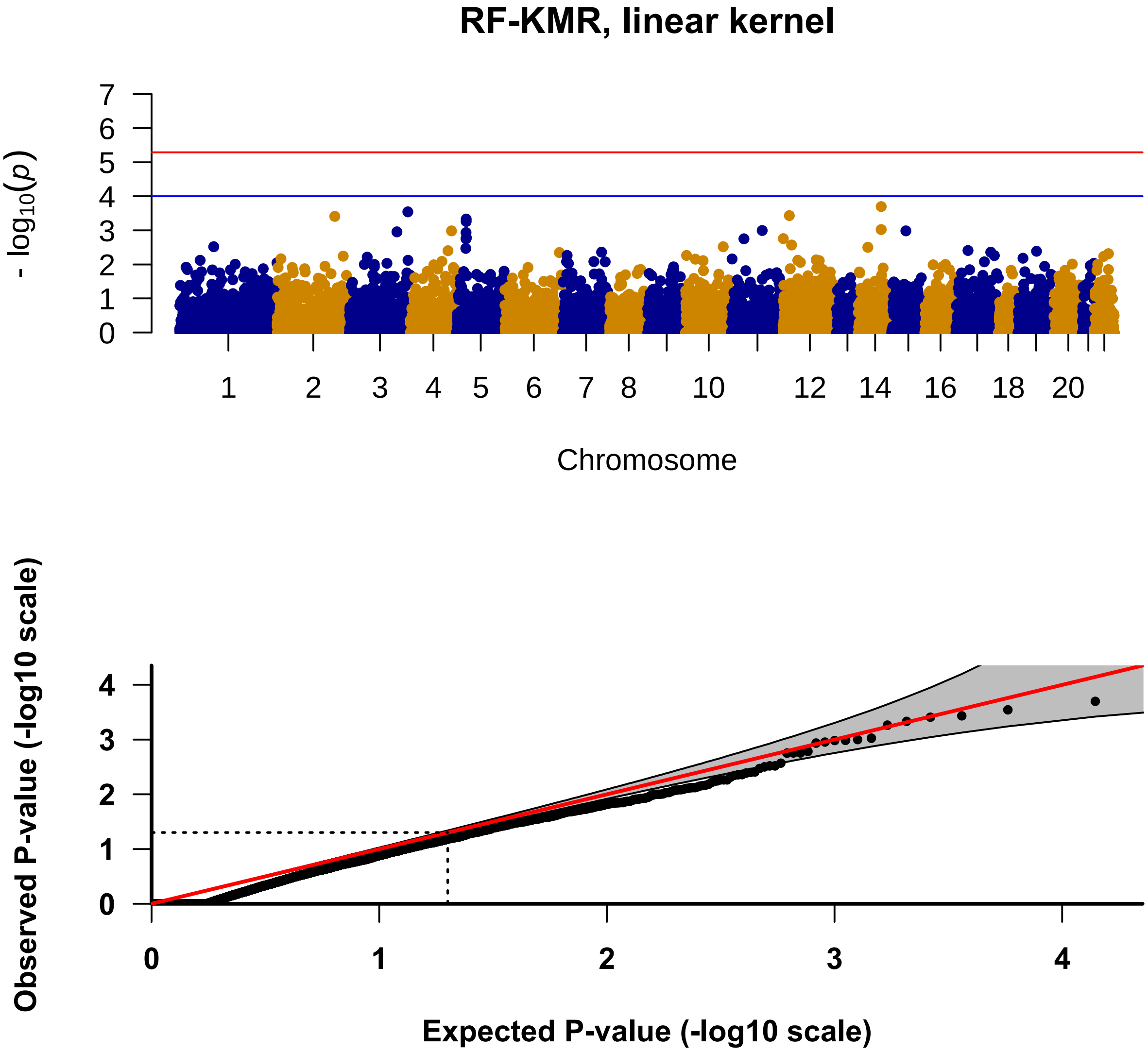
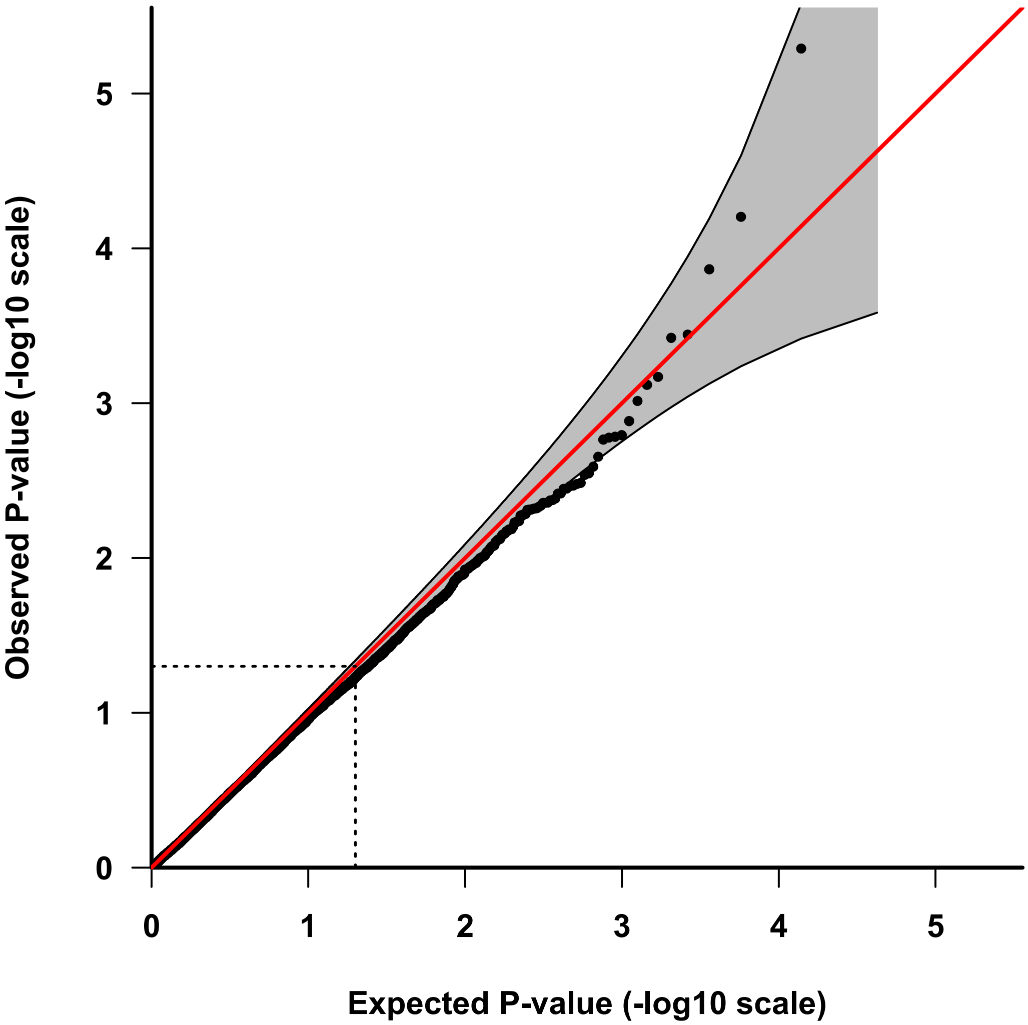
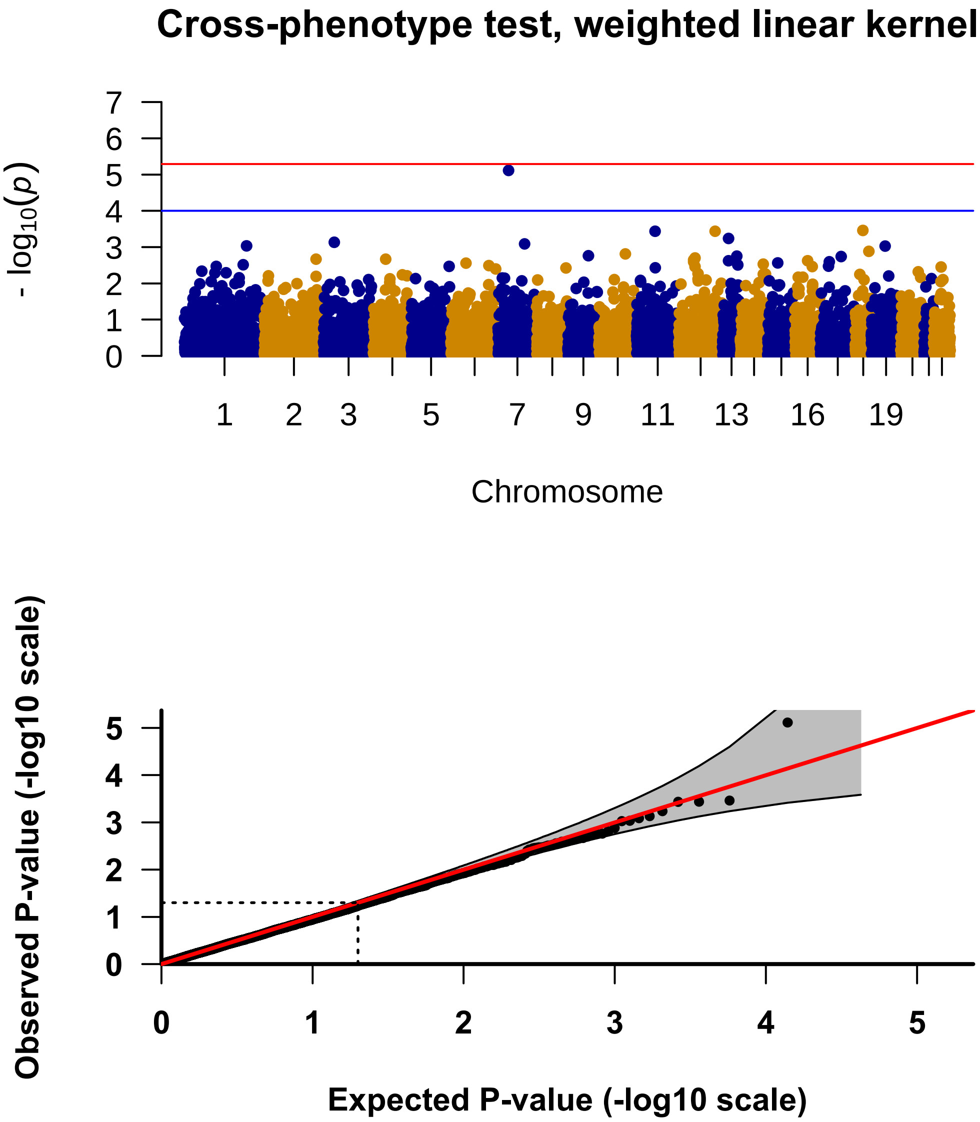


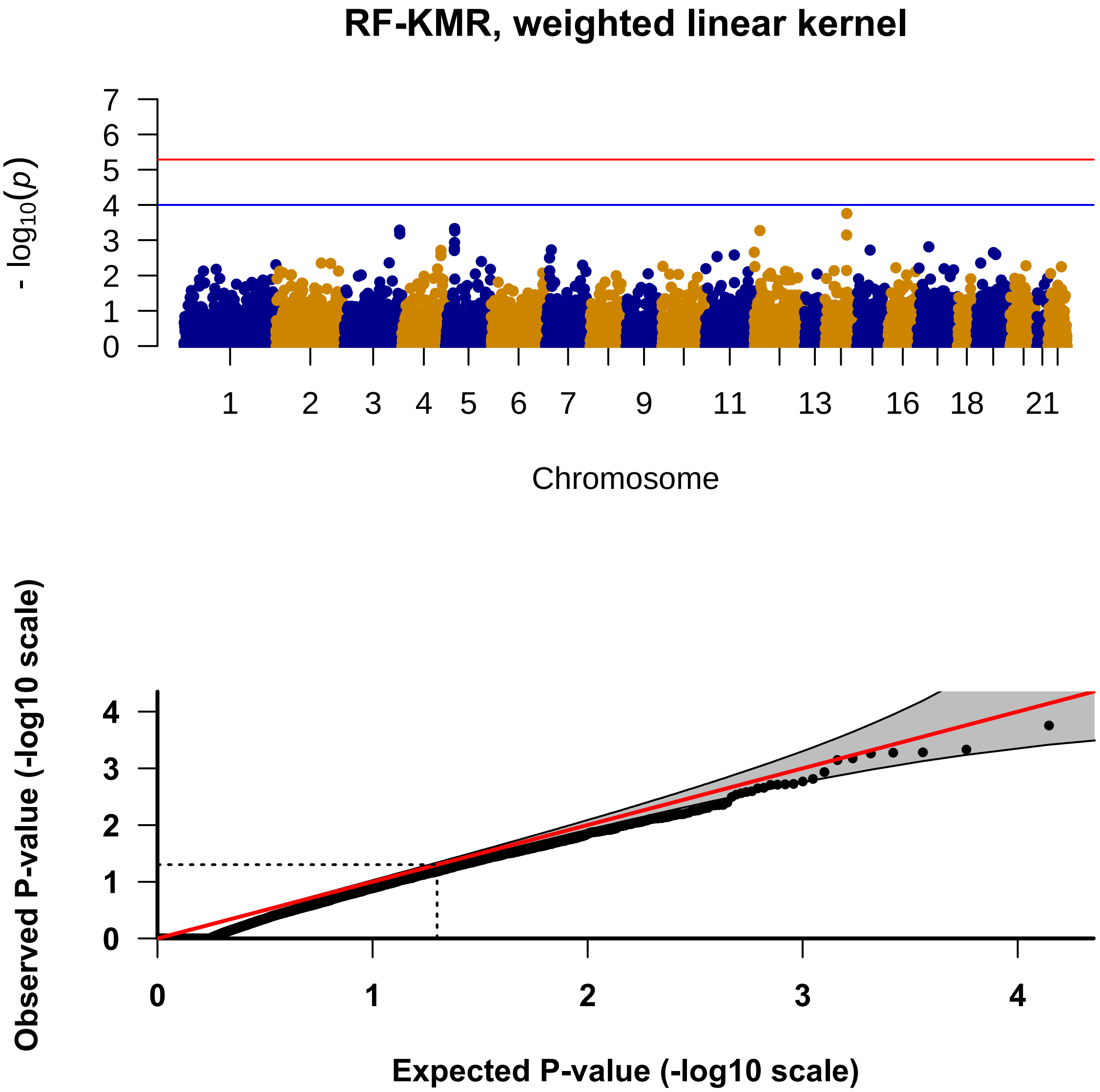
Figure 4.6 RF-KMR test on GoKinD data. Top: Manhattan plot for cross-phenotype test with linear kernel. Red line: genome-wide significance level. Blue line: suggestive level. Bottom: Quantile-Quantile Plo



Supplementary Figure 4.1. Quantile-Quantile Plot, test on the observed genotype, linear kernel.



Supplementary Figure 4.2. Cross-phenotype test on GoKinD data, weighted linear kernel. Top: Manhattan plot for cross-phenotype test with linear kernel. Red line: genome-wide significance level. Blue line: suggestive level. Bottom: Quantile-Quantile Plot



Supplementary Figure 4.3. RF-KMR test on GoKind data, weighted linear kernel. Top: Manhattan plot for cross-phenotype test with linear kernel. Red line: genome-wide significance level. Blue line: suggestive level. Bottom: Quantile-Quantile Plot

Chapter 5. Conclusions and Future Work

Next-generation sequencing studies continue to advance our knowledge of the genetic architecture of complex diseases and traits. Nevertheless, the complex and high-dimensional structure of next-generation sequencing data induces a variety of statistical challenges, particularly with regards to the analysis of rare trait-influencing variation. In particular, while a few methods have been proposed, methods for family-based studies of rare variants are greatly lacking in the literature. Our goal is to develop statistical methods that are suitable for studying related individuals and are both robust and powerful for detecting genes harboring rare trait-influencing variation. To achieve this goal, we developed the following three methods described below.

In the second chapter, we proposed a kernel-machine method for rare-variant sequencing studies in trios and nuclear families. Our method has the advantages of 1) substantially improved power comparing to the standard linear regression model 2) test statistics following an asymptotic distribution, which allows easy derivation of p-values (Davies 1980),3) insensitivity to situations where rare causal variants in a gene differ in their direction of effect, 4) robustness to population stratification through the integration of QTDT framework, and 5) two screening methods to boost power. We published our method in *Genetic Epidemiology* (Jiang, Conneely et al. 2014). The screening method proposed in our paper was subsequently adopted by other methodological research projects (Jiang, Ji et al. 2017). One limitation of our method in the second chapter is that our technique can only be applied to trios and nuclear families. As more studies tend to re-sequence subjects from larger pedigrees collected from previous linkage projects, we expanded our framework in the third chapter to create an extension of the method to permit analyses for general pedigrees that can accommodate arbitrary family size and structure..

By calculating the between-family component using all founders' information, the large-pedigree studies have the potential to improve power compared to nuclear family or trios.

One direction where we can extend the methods created in chapters 2 and 3 is to modify the approach to handle binary outcomes in family-based rare-variant sequencing studies. Binary traits, such as schizophrenia, also tend to have increased genetic load in families, and thus are appealing to study under family-based designs. However, very few existing methods for family-based studies can be applied to binary traits. Under the kernel machine framework, the logistic mixed model is difficult and computationally taxing to fit. Instead, we could use the generalized estimating equation (Wang, Lee et al. 2013) framework to rectify this issue. Another possible extension is to consider other weighting strategies for considering between-family and within-family information in our testing procedure. In particular, we can consider an idea similar to Mirea et al. (Mirea, Infante-Rivard et al. 2012), who adopted a weighting strategy where the between-family and within-family contributions to a test statistic are weighted by a test of population-stratification bias. We will explore these ideas in future work. We will also ensure these methods possess the same important features of the methods in chapters 2 and 3: robustness to population stratification and flexibility regarding direction of effect of causal variants in the tested region of interest.

In chapter 4, we developed a method for testing variants in genes that are pleiotropic and influence multiple diverse phenotypes using the kernel distance-covariance (KDC) framework. The KDC-based test statistics of our method are very easy to compute and follow a known asymptotic distribution. We applied our method both to simulated data and to GWAS data of multiple phenotypes from the GoKinD study of type 1 diabetes. Through simulation studies, we showed that our method is robust to population stratification and achieves higher statistical power than univariate analysis of each separate phenotype accounting for multiple testing. In real data analysis, we identified the VPS41 gene to be associated with diabetes related phenotypes (Systolic blood pressure, diastolic blood pressure, high-density lipoprotein, body mass index, which has not been reported before. In the future, we would like to extend our method to test the combined effect of rare and common variants within a gene or region of interest. The framework proposed by Ionita-Laza (Ionita-Laza, Lee et al. 2013) can be a possible direction for such extension. We plan to perform the KDC test on rare and common variants separately and then combine the test statistics through a weighted sum approach. Given that there is no cross-phenotype test to examine the combined effect of rare and common variants, we believe this approach can fill an interesting gap in the literature.

Reference

Abecasis, G., et al. (2000). "A general test of association for quantitative traits in nuclear families." The American Journal of Human Genetics **66**(1): 279-292.

Abecasis, G. R., et al. (2000). "A general test of association for quantitative traits in nuclear families." Am J Hum Genet **66**(1): 279-292.

Abecasis, G. R., et al. (2000). "Pedigree tests of transmission disequilibrium." European Journal of Human Genetics **8**(7).

Almasy, L., et al. (2014). Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. BMC proceedings, BioMed Central.

Bach, F. R. and M. I. Jordan (2002). "Kernel independent component analysis." Journal of machine learning research **3**(Jul): 1-48.

Broadaway, K. A., et al. (2016). "A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants." The American Journal of Human Genetics **98**(3): 525-540.

Cardon, L. R. and L. J. Palmer (2003). "Population stratification and spurious allelic association." Lancet **361**(9357): 598-604.

Chen, H., et al. (2013). "Sequence kernel association test for quantitative traits in family samples." Genetic epidemiology **37**(2): 196-204.

Chen, H., et al. (2013). "Sequence kernel association test for quantitative traits in family samples." Genet Epidemiol **37**(2): 196-204.

Cheung, C. Y., et al. (2014). "A statistical framework to guide sequencing choices in pedigrees." The American Journal of Human Genetics **94**(2): 257-267.

Cruceanu, C., et al. (2013). "Family-based exome-sequencing approach identifies rare susceptibility variants for lithium-responsive bipolar disorder 1." Genome **56**(10): 634-640.

Cruchaga, C., et al. (2012). "Rare variants in APP, PSEN1 and PSEN2 increase risk for AD in late-onset Alzheimer's disease families." PLoS One **7**(2): e31039.

Davies, R. B. (1980). "Algorithm AS 155: The Distribution of a Linear Combination of χ2 Random Variables." Journal of the Royal Statistical Society. Series C (Applied Statistics) **29**(3): 323-333.

Davies, R. B. (1980). "Algorithm AS 155: The distribution of a linear combination of χ 2 random variables." Journal of the Royal Statistical Society. Series C (Applied Statistics) **29**(3): 323-333.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S.L. (2002). Analysis of longitudinal data, Oxford University Press.

Do, R., et al. (2012). "Exome sequencing and complex disease: practical aspects of rare variant association studies." Hum Mol Genet **21**(R1): R1-9.

Epstein, M. and L. Kwee (2007). A powerful multilocus association test for quantitative traits. Genetic epidemiology, WILEY-LISS DIV JOHN WILEY & SONS INC, 111 RIVER ST, HOBOKEN, NJ 07030 USA.

Epstein, M. P., et al. (2012). "A permutation procedure to correct for confounders in case-control studies, including tests of rare variation." The American Journal of Human Genetics **91**(2): 215-223.

Epstein, M. P., et al. (2012). "A permutation procedure to correct for confounders in case-control studies, including tests of rare variation." Am J Hum Genet **91**(2): 215-223.

Fang, S., et al. (2012). "Two adaptive weighting methods to test for rare variant associations in family-based designs." Genet Epidemiol **36**(5): 499-507.

Fang, S., et al. (2013). "Detecting association of rare variants by testing an optimally weighted combination of variants for quantitative traits in general families." Annals of Human Genetics **77**(6): 524-534.

Fulker, D. W., et al. (1999). "Combined linkage and association sib-pair analysis for quantitative traits." The American Journal of Human Genetics **64**(1): 259-267.

Galesloot, T. E., et al. (2014). "A comparison of multivariate genome-wide association methods." PloS one **9**(4): e95923.

Gravel, S., et al. (2011). "Demographic history and rare allele sharing among human populations." Proceedings of the National Academy of Sciences **108**(29): 11983-11988.

Gretton, A., et al. (2007). A kernel statistical test of independence. Advances in neural information processing systems.

Guo, X., et al. (2013). "Genetic association test for multiple traits at gene level." Genetic epidemiology **37**(1): 122-129.

He, X. and J. Zhang (2006). "Toward a molecular understanding of pleiotropy." Genetics **173**(4): 1885-1891.

Hua, W. Y. and D. Ghosh (2015). "Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies." Biometrics **71**(3): 812-820.

Ionita-Laza, I., et al. (2013). "Family-based association tests for sequence data, and comparisons with population-based association tests." Eur J Hum Genet **21**(10): 1158-1162.

Ionita-Laza, I., et al. (2013). "Sequence kernel association tests for the combined effect of rare and common variants." The American Journal of Human Genetics **92**(6): 841-853.

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

Jiang, Y., et al. (2014). "Flexible and Robust Methods for Rare-Variant Testing of Quantitative Traits in Trios and Nuclear Families." Genetic epidemiology **38**(6): 542-551.

Jiang, Y., et al. (2013). "Assessing the impact of population stratification on association studies of rare variation." Human heredity **76**(1): 28-35.

Jiang, Y., et al. (2017). "Leveraging population information in family-based rare variant association analyses of quantitative traits." Genetic epidemiology **41**(2): 98-107.

Kang, H. M., et al. (2010). "Variance component model to account for sample structure in genome-wide association studies." Nat Genet **42**(4): 348-354.

Kocarnik, J. M. and S. M. Fullerton (2014). "Returning pleiotropic results from genetic testing to patients and research participants." JAMA **311**(8): 795-796.

Krumm, N., et al. (2013). "Transmission disequilibrium of small CNVs in simplex autism." Am J Hum Genet **93**(4): 595-606.

Kwee, L. C., et al. (2008). "A powerful and flexible multilocus association test for quantitative traits." The American Journal of Human Genetics **82**(2): 386-397.

Kwee, L. C., et al. (2008). "A powerful and flexible multilocus association test for quantitative traits." Am J Hum Genet **82**(2): 386-397.

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

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

Lee, S., et al. (2012). "Optimal tests for rare variant effects in sequencing association studies." Biostatistics **13**(4): 762-775.

Li, B. and S. M. Leal (2008). "Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data." Am J Hum Genet **83**(3): 311-321.

Lin, X. (1997). "Variance component testing in generalized linear models with random effects." Biometrika(84): 309-326.

Liu, D., et al. (2007). "Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models." Biometrics **63**(4): 1079-1088.

Liu, Q., et al. (2013). "Marbled inflation from population structure in gene-based association studies with rare variants." Genet Epidemiol **37**(3): 286-292.

Louis-Dit-Picard, H., et al. (2012). "KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron." Nature genetics **44**(4): 456-460.

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

Maity, A., et al. (2012). "Multivariate Phenotype Association Analysis by Marker-Set Kernel Machine Regression." Genetic epidemiology **36**(7): 686-695.

Manolio, T. A., et al. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

Marchini, J., et al. (2004). "The effects of human population structure on large genetic association studies." Nat Genet **36**(5): 512-517.

Mathieson, I. and G. McVean (2012). "Differential confounding of rare and common variants in spatially structured populations." Nature genetics **44**(3): 243-246.

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

Mirea, L., et al. (2012). "Strategies for genetic association analyses combining unrelated case-control individuals and family trios." American journal of epidemiology **176**(1): 70-79.

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

Mueller, P. W., et al. (2006). "Genetics of Kidneys in Diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes." Journal of the American Society of Nephrology **17**(7): 1782-1790.

Musunuru, K., et al. (2010). "Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia." New England Journal of Medicine **363**(23): 2220-2227.

Norton, N., et al. (2011). "Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy." Am J Hum Genet **88**(3): 273-282.

O'Reilly, P. F., et al. (2012). "MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS." PloS one **7**(5): e34861.

Ott, J., et al. (2011). "Family-based designs for genome-wide association studies." Nat Rev Genet **12**(7): 465-474.

Ott, J., et al. (2015). "Genetic linkage analysis in the age of whole-genome sequencing." Nature Reviews Genetics.

Oualkacha, K., et al. (2013). "Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness." Genet Epidemiol **37**(4): 366-376.

Paul, D. (2000). "A double-edged sword." Nature **405**(6786): 515-515.

Pezzolesi, M. G., et al. (2009). "Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes." Diabetes **58**(6): 1403-1410.

Plemel, R. L., et al. (2011). "Subunit organization and Rab interactions of Vps-C protein complexes that control endolysosomal membrane traffic." Mol Biol Cell **22**(8): 1353-1363.

Price, A. L., et al. (2010). "Pooled association tests for rare variants in exon-resequencing studies." The American Journal of Human Genetics **86**(6): 832-838.

Purcell, S., et al. (2005). "Parental phenotypes in family-based association analysis." Am J Hum Genet **76**(2): 249-259.

Ramagopalan, S. V., et al. (2011). "Rare variants in the CYP27B1 gene are associated with multiple sclerosis." Annals of Neurology **70**(6): 881-886.

Saad, M. and E. M. Wijsman (2014). "Power of Family-Based Association Designs to Detect Rare Variants in Large Pedigrees Using Imputed Genotypes." Genet Epidemiol **38**(1): 1-9.

Schaffner, S. F., et al. (2005). "Calibrating a coalescent simulation of human genome sequence variation." Genome research **15**(11): 1576-1583.

Schaid, D. J. (2010). "Genomic similarity and kernel methods II: methods for genomic information." Human heredity **70**(2): 132-140.

Schaid, D. J., et al. (2013). "Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data." Genet Epidemiol **37**(5): 409-418.

Schifano, E. D., et al. (2012). "SNP set association analysis for familial data." GenetEpidemiol **36**(8): 797-810.

Schifano, E. D., et al. (2013). "Genome-wide association analysis for multiple continuous secondary phenotypes." The American Journal of Human Genetics **92**(5): 744-759.

Schölkopf, B., et al. (1998). "Nonlinear component analysis as a kernel eigenvalue problem." Neural computation **10**(5): 1299-1319.

Serre, D., et al. (2008). "Correction of population stratification in large multi-ethnic association studies." PLoS One **3**(1): e1382.

Simpson, C. L., et al. (2011). Old lessons learned anew: family-based methods for detecting genes responsible for quantitative and qualitative traits in the Genetic Analysis Workshop 17 mini-exome sequence data. BMC proceedings, BioMed Central Ltd.

Sivakumaran, S., et al. (2011). "Abundant pleiotropy in human complex diseases and traits." The American Journal of Human Genetics **89**(5): 607-618.

Solovieff, N., et al. (2013). "Pleiotropy in complex traits: challenges and strategies." Nature Reviews Genetics **14**(7): 483-495.

Van Steen, K., et al. (2005). "Genomic screening and replication using the same data set in family-based association testing." Nat Genet **37**(7): 683-691.

Wang, J. L., et al. (2010). "TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing." Brain **133**(12): 3510-3518.

Wang, X., et al. (2013). "GEE-Based SNP Set Association Test for Continuous and Discrete Traits in Family-Based Association Studies." Genetic epidemiology **37**(8): 778-786.

Welter, D., et al. (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." Nucleic Acids Res **42**(Database issue): D1001-1006.

Wessel, J. and N. J. Schork (2006). "Generalized Genomic Distance–Based Regression Methodology for Multilocus Association Analysis." Am J Hum Genet **79**(5): 792-806.

Wijsman, E. M. (2012). "The role of large pedigrees in an era of high-throughput sequencing." Human genetics **131**(10): 1555-1563.

Wijsman, E. M. and C. I. Amos (1997). "Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions." Genetic epidemiology **14**(6): 719-735.

Wilson, A. F. and A. Ziegler (2011). "Lessons learned from Genetic Analysis Workshop 17: transitioning from genome-wide association studies to whole-genome statistical genetic analysis." Genetic epidemiology **35**(S1): S107-S114.

Wu, M. C., et al. (2010). "Powerful SNP-set analysis for case-control genome-wide association studies." The American Journal of Human Genetics **86**(6): 929-942.

Wu, M. C., et al. (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." Am J Hum Genet **89**(1): 82-93.

Yang, Q. and Y. Wang (2012). "Methods for analyzing multivariate phenotypes in genetic association studies." Journal of probability and statistics **2012**.

Zawistowski, M., et al. (2010). "Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes." The American Journal of Human Genetics **87**(5): 604-617.

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

Zhao, J. and A. Thalamuthu (2011). Gene-based multiple trait analysis for exome sequencing data. BMC proceedings, BioMed Central.

Zollner, S. (2012). "Sampling strategies for rare variant tests in case-control studies." Eur J Hum Genet **20**(10): 1085-1091.

Zöllner, S. (2012). "Sampling strategies for rare variant tests in case–control studies." European Journal of Human Genetics **20**(10): 1085-1091.