Design and Association Methods for Next-generation Sequencing Studies for Quantitative Traits

by

Shuang Feng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2015

Doctoral Committee:

Professor Gonçalo R. Abecasis, Chair
Professor Michael L. Boehnke
Professor Peter X. Song
Assistant Professor Cristen J. Willer

DEDICATION

To my husband Wei and our children Grace, Kevin, and Charles

ACKNOWLEDGEMENTS

I want to express my deepest gratitude to my advisor Gonçalo Abecasis for his mentorship during all these years. Without his guidance and trust, I would never be able to discover my potential of being able to do things that I would never have imagined being capable of. I could never be grateful enough of how his brilliant mind and his kindness have inspired and encouraged me. I would never learn enough from him. I want to express my gratefulness to my committee members Michael Boehnke, Peter Song, and Cristen Willer, for their relentless support in my dissertation research and study in graduate school.

I want to thank my fellow graduate students in Biostatistics department, especially class of 2015, Adrian, Sebanti, Tianshuang, Zhuqing, Kenghan, Yancy, and Sara. I also want to thank Mary Kate, Sean, Laura, Tara, and Irene for their help in various aspects for my research. You make the journey in graduate school not lonely.

I want to thank my parents for helping me caring my children and supporting me in my difficult times. Although they may not fully understand why I went back to graduate school after working full-time for four years and with children, I believe they are very proud of whatever I do and achieve.

I want to thank my children Grace and Kevin. I want to thank especially Grace for being such a kind and understanding child, for being self-motivated in school (Grace skipped second grade in Ann Arbor and was selected as a GATE student in California!), for being so keen and talented in many things, and for being so caring and loving baby brother Kevin, and for bringing our family so much love and happiness all these years. I want to thank Kevin for being such an adorable and cute baby and bringing so much happiness to our busy life.

Most importantly, I want to thank my husband for his love on me and our children, for his being such a responsible husband and father, for his encouragement and support in my difficult times, for his endurance of my late-night-coding hobby during graduate school, for his knowing me for 21 years and marrying me for 12 years (till 2015), and for the many more happy years that we are going to spend together.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Advances in exome sequencing and the development of exome genotyping arrays are enabling explorations of association between rare coding variants and complex traits using sequencing-based GWAS. However, the cost of sequencing remains high, optimal study design for sequencing-based association studies is an open question, powerful association methods and software to detect trait-associated rare and low-frequency variants are in great need. Containing 5% of information in human genome sequence, chromosome X analysis has been largely neglected in routine GWAS analysis. In this dissertation, I focus on three topics:

First, I describe a computationally efficient approach to re-construct gene-level association test statistics from single-variant summary statistics and their covariance matrices for single studies and meta-analyses. By simulation and real data examples, I evaluate our methods under the null, investigate scenarios when family samples have larger power than population samples, compare power of different types of gene-level tests under various trait-generating models, and demonstrate the usage of our methods and the C++ software, RAREMETAL, by meta-analyzing SardiNIA and HUNT data on lipids levels.

Second, I describe a variance component approach and a series of gene-level tests for X-linked rare variants analysis. By simulations, I demonstrate that our methods are well

controlled under the null. I evaluate power to detect an autosomal or X-linked gene of same effect size, and investigate the effect of sex ratio in a sample to power of detecting an X-linked gene. Finally I demonstrate usage of our method and the C++ software by analyzing various quantitative traits measured in the SardiNIA study and report detected X-linked variants and genes.

Third, I describe a novel likelihood-based approach and the C++ software, RAREFY, to prioritize samples that are more likely to be carriers of trait-associated variants in a sample, with limited budget. I first describe the statistical method for small pedigrees and then describe an MCMC approach to make our method computationally feasible for large pedigrees. By simulations and real data analysis, I compare our approach with other methods in both trait-associated allele discovery power and association power, and demonstrate the usage of our method on pedigrees from the SardiNIA study.

CHAPTER 1: INTRODUCTION

Mapping Human Complex Traits

Focus of gene mapping of human complex traits migrated from linkage studies to association studies gradually since the end of the 20th century. As pointed out by [Risch and Merikangas 1996], association studies can have greater power than linkage studies but were limited by the fact that not many polymorphisms or genes were identified at that time. The number of markers that were available for analysis was usually in the tens and sample size was in the hundreds.

Advances in genotyping technology and rapidly reduced genotyping cost in the beginning of the 21st century have facilitated detecting a large amount of polymorphisms across the entire human genome and brought a plethora of discoveries through genome-wide association studies (GWAS) for various human complex diseases and traits [Teslovich et al. 2010; Willer et al. 2008]. Genotyping arrays allow scientists to analyze association of variants that are in linkage disequilibrium with causal variants, instead of analyzing markers that might be several cM away from the causal gene on the same chromosome in linkage studies. However, , because not enough features could be captured on the chip, polymorphisms studied in GWAS were usually relatively common in frequency, which has been shown to contribute to a very small proportion of human DNA variations according to the observations from 1000 Genome Project [Abecasis et al. 2010]. The fact that GWAS findings altogether were not able to fully explain the trait variance brought

up the well-known "missing heritability" question. A natural conjecture for a possible solution of this question was that rare and low-frequency variants with large effects might exist and could explain the missing heritability, but we were simply not able to detect them.

Advances in Sequencing Studies

While array-basedGWAS continued to succeed, sequencing technology has been improving at a fast speed and sequencing cost has been decreasing rapidly. In 2014, a decade after the completion of the first GWAS, whole-genome sequencing cost reached the $1,000 per genome milestone. Unlike in array-based GWAS, sequencing makes it possible to analyze causal genes and variants directly instead of studying their linkage disequilibrium proxies - the common variants that are in linkage disequilibrium with them. Many sequencing studies have been conducted or are on-going [Lange et al. 2014; T2D-GENES-Consortium,In Preparation].

Sequencing also allows discoveries of rare and low-frequency variants with moderate to large effects which are expected to explain at least part of the "missing heritability" mystery. Although, at present, there is no sequencing study that is able to show that rare variant discoveries could explain the missing heritability from GWAS findings fully for any trait, thorough investigations of rare and low-frequency variants are expected to bring biological insights to biology of human diseases and traits because rare variants are more likely to be functional [Nelson et al. 2012]. Besides whole-genome sequencing, exome sequencing and exome chip arrays are also cost-effective strategies for rare and low-

frequency discoveries and significant amount of successful findings have been reported [Crosby et al. 2014; Lange et al. 2014].

Sequencing studies can also provide valuable source to generate imputation reference panels to increase imputation accuracy for rare and common variants. Imputation using reference panels generated from 1000 Genome sequencing study has become routine analysis in GWAS. Sequencing a proportion of samples from a cohort to build an enriched reference panel with disease-causing rare mutations together with the currently existing reference panel followed by imputation to large well-phenotyped cohort can greatly enlarge power to detect trait-associated or disease-causal genes [Hoffmann et al. 2015]. The Haplotype Reference Consortium (HRC) [REF] creates a large reference panel of human haplotypes combining whole-genome sequencing data from multiple cohorts, enabling imputation of large amount of rare and low-frequency variants to enlarge GWAS power.

Challenges in Sequencing-based GWAS

Challenges are non-negligible for design and analysis of sequencing-based genome-wide association studies. First, optimal study design for sequencing-based GWAS is an open question. Family samples were essential for linkage studies because transmission patterns which are the core to detect linkage signal, can be tracked or inferred in pedigrees. Array-based GWAS largely used population samples, because unrelated samples naturally have large power to detect common variants in association than family samples. However, in sequencing-based GWAS, rare and low-frequency variants are the focus for association analysis, and sampling unrelated individuals from a population requires very

large sample size to be able to capture enough rare alleles for enough power. Family samples sometimes can capture more than average copies of rare alleles due to the "Jackpot" effect where multiple copies of a rare allele can be observed in a single pedigree. But how this could affect association power remains an open question. What's more, sequencing cost remains high. Novel methods for cost-efficient study design of sequencing studies are in great need.

Second, detecting the associations of rare and low-frequency variants that contribute to the majority of polymorphisms from sequencing data has extremely limited power unless there are large enough samples to be sequenced or large enough number of rare alleles captured. One popular strategy is to aggregate rare variants within a gene or a region to bring a synergy of information to enlarge power. Many different statistical methods have been published and they can be summarized into two types of gene-level association methods based on whether genetic effect of a gene is modeled as a fixed effect, for example, in burden [Li and Leal 2008; Madsen and Browning 2009] and Variable Threshold [Lin and Tang 2011; Price et al. 2010] tests, or random, for example, Sequence Kernel Association Test (SKAT) [Wu et al. 2011] and SKAT-O [Lee et al. 2012a]. These gene-level tests have certain advantages and disadvantages for different disease models. For example, burden type tests are more powerful when causal variant counts in a gene is large and all causal variants have effect sizes of the same direction, yet loses power quickly when causal variants are bi-directional; kernel-based variance component tests, such as SKAT, are most powerful when causal variant count is small or causal variants have opposite directions in effect sizes. The other popular strategy is Meta-analysis,

which has been extremely successful in array-based GWAS studying common variants [Scott et al. 2007; Willer et al. 2010]. Meta-analysis naturally enlarges power by increasing sample sizes without sharing raw data. However, powerful meta-analysis association methods for sequencing data for family samples are sparse. Many association methods for quantitative traits to detect single common variant associations in nuclear families and sib-pairs [Abecasis et al. 2000; Abecasis et al. 2001b; Laird et al. 2000], and in general pedigrees [Chen and Abecasis 2007] have been published. But gene-level tests and meta-analysis methods for rare variant associations in families are under development.

Third, chromosome X association analyses have been largely neglected in array-based GWAS, although containing 5% of human DNA sequences. X-linked QTL linkage analysis methods have been extensively studies and implemented in tools that are widely used [Abecasis 2002; Abecasis et al. 2000; Almasy and Blangero 1998; Lange and Sobel 2006]. X-linked single variant association tests for quantitative traits in unrelated and related samples have also been studied [Abecasis 2002; Clayton 2008; Clayton 2009; Zhang et al. 2009]. However, gene-level association methods and meta-analysis methods for X-linked rare and low-frequency variants are in great need.

Outline of this Thesis

In this dissertation, I focus on three topics related to design and association methods for sequencing data analysis for quantitative traits. First, I describe the gene-level association and meta-analysis methods for sequencing data in family and population samples [Feng et al. 2015]. Second, I address the association methods for X-linked rare and low-

frequency variants in family and population samples. Third, I describe a novel likelihood-based method to prioritize samples that are more likely to carry trait-associated rare variants.

In chapter 2, I investigate the advantages and disadvantages of family and population samples in modern genetic association studies, especially sequencing-based GWAS. I describe scenarios when family samples have more power than population sample using simulation. I then propose efficient gene-level association methods for single studies and for meta-analysis of family and population samples. By simulations, I seek to 1) demonstrate that our methods are well calibrated under the null by calculating type I errors and 2) compare power of different gene-level association tests under various trait-generating model and 3) compare power and computational performance of our method and software with other published methods and tool. Finally, using SardiNIA and HUNT exome chip data, I demonstrate the usage of our methods in meta-analysis by finding confirmed trait-associated genes for blood lipid traits.

In chapter 3, I describe our statistical approach to analyze X-linked rare variants. I describe a variance component model to properly handle relatedness and cryptic relatedness and population structure in a sample. By simulations, I demonstrate that 1) our methods are under control under the null and 2) there is larger power to detect a X-linked gene than an autosomal gene with the same effect when complete X-inactivation is assumed and then I further evaluate the relationship between power and proportion of females in a sample. Finally, using SardiNIA quantitative traits and exome chip data, I

demonstrate the usage of our method and tool and report associated X-linked genes and rare variants to some of the quantitative traits measured in SardiNIA sample.

In chapter 4, I describe a novel likelihood-based approach to select samples that are more likely to carry trait associated rare variants in a currently existing sample, with limited sequencing cost. I first describe the statistical method for small pedigrees and then describe an MCMC approach for large pedigrees to make our method computationally feasible. By simulations, I compare our approach with methods that select phenotypic extremes by evaluating both trait-associated allele discovery power and association power, and I demonstrate that our method is not affected by the choice of prior values of frequency and effect size. Finally, using SardiNIA data, I demonstrate the usage of our method on large pedigrees (as many as ~1,200 individuals per family) and show that our method has larger discovery power than the competing method which considers only phenotypic extremes.

In chapter 5, I summarize my work and propose possible interesting topics in design and analysis methods for sequencing-based studies.

CHAPTER 2: STUDY DESIGN AND ASSOCIATION AND META-ANALYSIS METHODS IN FAMILIES

Introduction

Variants of functional consequence, including non-synonymous, splice altering, and protein truncating variants, usually segregate at very low frequency in human populations [Abecasis et al. 2010; Abecasis et al. 2012; Marth et al. 2011; Nelson et al. 2012]. Recent advances in exome sequencing and the development of exome genotyping arrays are enabling explorations of their contributions to complex disease [Kiezun et al. 2012].

Association of rare variants with disease will bring biological insights about disease processes, but standard variant-by-variant association tests lack power when applied to these variants unless sample sizes are very large. Our work builds upon three strategies to increase the power of rare variant association studies: grouping variants by gene or functional unit, combining results across many studies through meta-analysis, and analysis of family samples.

Grouping rare variants by gene or functional unit [Li and Leal 2008], whether with weights [Madsen and Browning 2009] or without [Morris and Zeggini 2010], is now a popular strategy for rare variant association analysis [Lee et al. 2012a; Lee et al. 2012b; Lin and Tang 2011; Price et al. 2010; Wu et al. 2011]. The approach assumes

that rare variants in the same gene or functional unit have similar functional consequences. When the assumption is correct and rare variants in a region are analyzed together, association signals will be stronger than when evaluating variants individually.

A second strategy to increase power is meta-analysis, which increases sample size and provides a practical approach to difficulties in data-sharing and concerns about heterogeneity [Lin and Zeng 2010; Willer et al. 2010]. Meta-analysis of single variants has been key in establishing association between common variants and complex diseases [Scott et al. 2007; Willer et al. 2010]. Meta-analysis methods for rare variant association tests have now been proposed, although these initial proposals and their implementations have generally focused on samples of unrelated individuals [Lee et al. 2013; Liu et al. 2014; Tang and Lin 2013].

Finally, a third strategy is to study samples of closely related individuals, increasing the odds that multiple copies of each rare variant are observed. Family samples are key in studies of Mendelian disorders but can also have advantages for studies of complex traits [Laird and Lange 2006, 2008; Ott et al. 2011]. For example, they can be more robust to population stratification (which may be more acute in rare variant association studies [Gravel et al. 2011]), allow checks for genotyping errors, improving data quality [Abecasis et al. 2001a; Abecasis et al. 2002][Abecasis et al. 2001; Abecasis et al. 2002] and can be enriched for variants of large effect by focusing on families with multiple individuals with extreme phenotypes. Early tests

for family based association [Abecasis et al. 2000; Laird et al. 2000; Laird and Lange 2008] focused on analysis of transmission disequilibrium, but newer tests rely on variance component models [Chen and Abecasis 2007; Kang et al. 2010] to account for stratification, resulting in tests of association that are typically more powerful [Chen and Abecasis 2007]. Our work also builds on computational enhancements in methods for variance component analysis, which have now been extended to samples of unrelated individuals (using empirical kinship matrices, estimated from genotype data) [Kang et al. 2010; Lippert et al. 2011; Zhou and Stephens 2012].

Here, we describe family-based association tests for rare variants that allow analysis of quantitative traits, with or without covariates, and show how these tests can be applied in meta-analysis settings. Our methods are based on the insight that gene-level test statistics can be constructed from single variant score statistics and estimates of the covariance between those [Liu et al. 2014]. We first analyze single variants using efficient computational algorithms for evaluation of variance component models [Lippert et al. 2011]. We then develop family-based burden (weighted and un-weighted), sequence-kernel association (SKAT), and variable frequency threshold (VT) tests. Using simulation we show that type I error is well controlled and compare different testing approaches. As expected, SKAT tests are more powerful when the fraction of associated variants in each gene is small or associated rare variants have opposite directions of effect; VT tests are more robust to the choice of allele frequency threshold for grouping variants. Our analysis of exome chip genotypes and HDL level data from the HUNT and SardiNIA studies

shows that our methods are well calibrated and powerful enough to identify several signals at lipid associated loci.

There has been much recent work focused on extending gene-level association tests to families. Examples include various family-based burden tests [De et al. 2013; Saad and Wijsman 2014; Schaid et al. 2013] and variance component based tests [Chen et al. 2013; Ionita-Laza et al. 2013; Saad and Wijsman 2014; Schaid et al. 2013; Schifano et al. 2012; Svishcheva et al. 2014]. A key difference in our implementation, compared to previous work is that we construct our gene-level statistics using single-variant statistics as input. This allows us to quickly re-evaluate gene-level statistics when gene definitions or variant masks change, makes it practical to implement variable frequency-threshold based tests, and facilitates meta-analyses. To ensure computational efficiency in genome-wide analyses, our implementation uses a score-test that requires fitting a maximum likelihood model only once, rather than a Wald-test that would require it for every gene [Saad and Wijsman 2014]Madsen and Browning 2009]. We also focused on methods that could accommodate a diverse mix of family structures or even samples that include both families and unrelated individuals. This is in contrast to transmission-based tests [De et al. 2013; Ionita-Laza et al. 2013] that are limited to simpler family structures and cannot account for cryptic relatedness. As usual, we expect transmission based tests may provide greater protection against stratification – but at the cost of greatly reduced power.

We characterize settings where family studies can provide greater power to detect rare variants with moderate to large phenotypic consequences than studies of unrelated individuals. In studies of unselected samples, this is due to a "Jackpot" effect, where multiple copies of a extremely rare allele can be observed in a single pedigree. While for each locus the expected number of rare alleles will be the same in a family sample or an unrelated sample of same size, family samples are much more likely to exceed this expectation by a large amount. Our simulations show that this difference can have a large impact on power. All the methods described here are implemented in freely available C++ code and tools.

Methods

In this section, we first describe a variance component model to handle familial relationships. Then, we describe how single variant association statistics and their covariance matrices can be calculated and how gene-level association tests can be constructed. Next, we describe meta-analytic approaches for both single variant and gene-level association tests. Finally, we discuss the computational cost of our proposed approach and provide practical suggestions to improve computational performance.

Modeling Relatedness

In a sample of n individuals, we model the observed phenotype vector ($\mathbf{y}$) as a sum of covariate effects (specified by a design matrix $\mathbf{X}$ and a vector of covariate

effects $\boldsymbol{\beta}$), additive genetic effects (modeled in vector $\mathbf{g}$) and non-shared environmental effects (modeled in vector $\boldsymbol{\epsilon}$). Thus:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{g} + \boldsymbol{\epsilon}. \qquad \text{(Equation 1)}$$

We assume that genetic effects are normally distributed, with mean $\mathbf{0}$ and covariance $2\sigma_g^2\mathbf{K}$ where the matrix $\mathbf{K}$ summarizes kinship coefficients [Lange 1997] between sampled individuals and σ_g^2 is a positive scalar describing the genetic contribution to the overall variance. We assume that non-shared environmental effects are normally distributed with mean $\mathbf{0}$ and covariance $\mathbf{I}\sigma_e^2$, where $\mathbf{I}$ is the identity matrix.

To estimate $\mathbf{K}$, we either use recorded pedigree structure to define $\widehat{\mathbf{K}}$ following the method described in [Lange 1997] or else use the Balding-Nicols empirical estimator [Astle and Balding 2009], which uses observed genotypes to estimate kinship as $\widehat{\mathbf{K}} = \frac{1}{v}\sum_{i=1}^{v} \frac{(\mathbf{G_i} - 2f_i\mathbf{1})(\mathbf{G_i} - 2f_i\mathbf{1})^{\mathbf{T}}}{4f_i(1-f_i)}$ (here, v is the count of variants, $\mathbf{G_i}$ is a genotype vector where each element encodes the number of observed minor alleles in a particular individual, and f_i is the estimated allele frequency for the i^{th} variant). Model parameters $\hat{\beta}$, $\widehat{\sigma_g^2}$ and $\widehat{\sigma_e^2}$, are estimated using maximum likelihood and the efficient algorithm described in Lippert et al. [Lippert et al. 2011]. For convenience, let the estimated covariance matrix of $\mathbf{y}$ be $\widehat{\boldsymbol{\Omega}} = 2\widehat{\sigma_g^2}\widehat{\mathbf{K}} + \widehat{\sigma_e^2}\mathbf{I}$.

Single-variant Association Tests and Summary Statistics

Since our gene level association tests will build on single-variant test statistics [Chen and Abecasis 2007], we will first describe single variant test statistics and their corresponding variance-covariance matrix.

Consider the model

$$\mathbf{y} = \mathbf{X}\beta + \gamma_i(\mathbf{G_i} - \overline{\mathbf{G}}_\mathbf{i}) + \mathbf{g} + \varepsilon .$$

This model is a refinement of equation (1) above, adding a scalar parameter $\gamma_\mathbf{i}$ to measure the additive genetic effect of the *i*th variant. As usual [Lange 1997], the score statistic for testing $\mathbf{H}_0 : \gamma_i = 0$ is

$$U_i = (\mathbf{G_i} - \bar{\mathbf{G}}_\mathbf{i})^T \hat{\boldsymbol{\Omega}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta})$$

And the variance-covariance matrix of these statistics is:

$$\mathbf{V} = (\mathbf{G} - \bar{\mathbf{G}})^T (\hat{\boldsymbol{\Omega}}^{-1} - \hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}(\mathbf{X}^T\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\Omega}}^{-1})(\mathbf{G} - \bar{\mathbf{G}}).$$

Under the null, test statistics $T_i = \frac{U_i^2}{V_{ii}}$ are asymptotically distributed as chi-squared with one degree of freedom.

Gene-level Association Tests for Family Samples

Using single variant statistics U_i and their variance-covariance matrix **V**, we are now ready to construct a variety of gene-level association test statistics that combine information across variants.

The simplest statistic for a burden test is to estimate the average genetic effect across a series of variants satisfying certain functional (for example, non-

synonymous or protein truncating variants) and frequency criteria (for example, allele frequency <.05). Then the rare variant burden for each individual can be defined as a weighted sum of allele counts for variants satisfying these criteria. Abstractly, we define the rare variant burden as $(\mathbf{G}-\bar{\mathbf{G}})\mathbf{w}$, where $\mathbf{w}=(w_1,w_2,\ldots,w_m)^T$ is a vector of weights for each of the m variants in the gene. A regression parameter measuring the average effect of each variant can be estimated using the model:

$$\mathbf{y}=\mathbf{X}\beta+\gamma(\mathbf{G}-\bar{\mathbf{G}})\mathbf{w}+\mathbf{g}+\varepsilon.$$

To test the null hypothesis $\gamma=0$, we use a score statistic, expressed as a function of single variant statistics $\mathbf{w}^T\mathbf{U}$ with variance $\mathbf{w}^T\mathbf{V}\mathbf{w}$.

Then the burden test statistic $T_{burden}=\frac{\mathbf{w}^T\mathbf{U}}{\sqrt{\mathbf{w}^T\mathbf{V}\mathbf{w}}}$ is asymptotically normal with mean zero and variance one.

Variable Threshold Tests for Family Samples

The simplest burden tests will be effective when appropriate frequency thresholds and functional annotation are used to select functional variants for analysis. However, this is challenging to do, because the optimal frequency thresholds will vary by gene and by phenotype [Lange et al. 2014]. One possibility is to define a test statistic that considers many alternative frequency thresholds [Lin and Tang 2011; Price et al. 2010].

Following the suggestions of Price et al. 2010 and Lin et al. 2011, we will define the variable threshold test statistic as the maximal absolute value of burden test statistics across all possible frequency thresholds, $T_{VT} = \max_F |T_{burden F}|$, where $T_{burden F} = \frac{\phi_F^T U}{\sqrt{\phi_F^T V \phi_F}}$ is the burden test statistic calculated with frequency threshold *F* and ϕ_F is a vector of 0s and 1s indicating whether a variant has allele frequency below *F*. Burden statistics calculated using different frequency thresholds jointly follow a multivariate normal distribution with mean **0**, and variance-covariance matrix $\psi_{ij} = \frac{\phi_i^T V \phi_j}{\sqrt{\phi_i^T V \phi_i}\sqrt{\phi_j^T V \phi_j}}$ [Lin and Tang 2011]P-values can be evaluated using the cumulative density function of this multivariate normal distribution [Genz 1992].

Sequence Kernel Association Tests

Another refinement is to use a test statistic that allows for variants in the same gene to modify the phenotype in opposite directions [Chen et al. 2013; Ionita-Laza et al. 2013; Wu et al. 2011; Yan et al. 2014]. For example, in some genes [Abifadel et al. 2003], both gain-of-function and loss-of-function alleles have been described and these signals might cancel each other in a standard burden analysis. The model for this type of test is

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\gamma}(\mathbf{G} - \bar{\mathbf{G}}) + \mathbf{g} + \boldsymbol{\varepsilon},$$

In this alternative model, the single variant effects γ_i are assumed to follow a shared distribution, with mean 0 and variance tw_i. We test the null hypothesis of no

association using the statistic $T_{SKAT} = \mathbf{U}^T\mathbf{W}\mathbf{U}$ to evaluate whether τ is nonzero [Chen et al. 2013; Wu et al. 2011]. As usual, $\mathbf{W} = \mathrm{diag}(w_1, w_2, \ldots, w_m)$ is a diagonal matrix indicating the weight of each variant. T_{SKAT} is distributed as a mixture chi-squared with weights $\lambda_1, \lambda_2, \ldots, \lambda_n$ corresponding to the eigenvalues of $\mathbf{V}^{\frac{1}{2}}\mathbf{W}\mathbf{V}^{\frac{1}{2}}$, and the $\chi_1^2(i)$ correspond to independently distributed chi-squared variables, each with 1 degree of freedom [Wu et al. 2011]. P-values can be approximated using the Davies algorithm [Davies 1980] or a moment matching algorithm [Liu et al. 2009].

Meta-Analysis

Since we derived all the statistics above from single variant score statistics and their covariance matrix, our approach can be readily extended to meta-analyses. We first define the overall single variant score statistics and their variance-covariance matrix as $U_{meta_i} = \sum_{k=1}^{s} U_{ik}$ and $V_{meta_{ij}} = \sum_{k=1}^{s} V_{ij,k}$, where U_{ik} and $V_{ij,k}$ are the single variant score statistic and variance-covariance matrix components from study k and s is the total number of studies. Whenever variant i is unobserved in study k, we set $U_{ik} = 0$ and $V_{ij,k} = 0$ for all *j*. Next, we simply calculate burden, VT and SKAT meta-analysis statistics using the formulae above.

Computational Efficiency

Since we rely on score statistics and their covariance, we only need to fit the linear mixed model once under the null hypothesis. Fitting parameters for this null mixed model is a major part of the computational cost of our approach. Standard EM or

Newton–Raphson methods require calculating the inverse of the covariance matrix in each iteration – with time complexity $O(n^3)$, too costly for large datasets. Instead, we used the computationally efficient algorithm described in [Lippert et al. 2011] to estimate the variance components and fixed effects under the null (**Equation 1**). The algorithm begins with a one-time singular value decomposition (SVD) of the relationship matrix $\hat{\mathbf{K}}$, a step which has time complexity $O(n^3)$. The results of this decomposition are used in a factorization that transforms the phenotype vector and design matrix so that transformed phenotypes are identically and independently distributed. This second step has time complexity $O(n^2)$. After transformation, the cost of updating the log likelihood becomes linear with respect to sample size n (instead of $O(n^3)$ using the standard approach). Calculating the score statistics and their covariance for all single variants simply requires a transformation of genotypes and has time complexity $O(mn^2)$ for a dataset with m variants. In reality, we calculate covariance of score statistics from markers within a sliding window. For large samples, calculating the SVD of $\hat{\mathbf{K}}$ is the computationally most expensive step. A similar idea with comparable computational efficiency has also been described in Zhou and Stephens (2012]. Both ideas build upon the algorithm described by Kang et al. and implemented in EMMAX [Kang et al. 2010].

When variants are grouped in gene-level tests, the computational cost of calculating the combined test statistics is small after single variants have been analyzed. Obtaining p-values corresponding to these statistics, especially for SKAT and VT analyses, can still be challenging when the number of rare variants in a gene is large.

To speed up this step, we used computationally efficient algorithms to evaluate the multivariate normal probabilities [Genz 1992] and the mixture chi-squared distribution [Davies 1980].

Simulation

We carried out a series of simulations to evaluate the type I error and power of our method. We first simulated a set of 1000 base-pair sequences, which is close to the length of an average protein coding sequence in humans, using the coalescent (as implemented in the program ms [Hudson 2002]) and a demographic model calibrated to mimic European population history [Adams and Hudson 2004; Novembre et al. 2008]. We then carried out gene-dropping simulations [Abecasis et al. 2002] using these simulated sequences as founder haplotypes that were propagated through various pedigree structures (**Figure 2.1**).

To evaluate power, we assigned a fraction of variants below a desired frequency threshold (<0.01 in simulations unless addressed otherwise) as causal. Typically, we assigned minor alleles at causal variants to all have effects in the same direction but, in some cases, a fraction of causal minor alleles were assigned effects in the opposite direction. When assigning effect sizes to causal variants, we considered two trait-generating models - an equal variance model (where the effect size for each variant is proportional to $\frac{1}{\sqrt{p(1-p)}}$, a function of the allele frequency *p* that ensures each causal variant explains the same amount of trait variance) and an equal effect-size

model (where the effect size is the same for all causal variants, irrespective of allele frequency). In the equal effect size model, relatively common variants explain a larger amount of the variance; while in the equal variance model, rarer variants have larger effect sizes (See **Figure S2.1** for demonstration). Genetic effects were set so that the total variance explained by each gene (h^2_{gene}) was in the 0.1-2% range. Empirical power was calculated using 10,000 simulations for each parameter combination. We used α=1x10^{-8} for single variant association power and α=2.5x10^{-6} for gene-level association power, following the consensus of significance level used in GWAS with Bonferroni correction. Type I error rate for gene-level tests was estimated using 5,000,000 simulations. To compare studies of families and unrelated individuals, we held the number of genotyped (or sequenced) individuals constant and compared our power to detect associated variants in studies using different sampling units. In simulations and following association analysis, kinship matrices estimated from pedigree were used to fit the null linear mixed model.

SardiNIA and HUNT Samples Description

To demonstrate usage of our methods in real data analysis, we used exome chip data from the HUNT [Holmen et al. 2014a; Holmen et al. 2014b] and SardiNIA [Giorgio et al. 2014; Pilia et al. 2006] studies, which genotyped 5,803 and 6,602 individuals, respectively. Here, we analyze HDL, adjusted for age and sex (**Table S2.3**). Genotypes were called using the Illumina GenCall algorithm in combination with zCall V2.2. Detailed QC procedures can be found in Holmen et al. [Holmen et al.

2014a] for the HUNT study and Pistis et al. [Giorgio et al. 2014] for the SardiNIA study.

Results

Type I Error Rate

To evaluate type I error rate, we simulated family samples of 1,000 or 5,000 individuals with with families of 3 generation pedigrees with 10 (Pedigree10) or 50 (Pedigree50) individuals (see **Figure 2.1** for details). Within each gene, variants with frequency <.01 were grouped for analysis. Each type I error estimate summarizes results from five-million simulations. **Table S2.1** shows that the type I error of our gene-level association tests is well controlled for a variety of pedigree structures. Empirical error rates are a little below nominal levels when sample sizes are small (N=1,000), but approach nominal significance as sample size increases (N=5,000).

Power of Different Rare Variant Association Tests

Next, we evaluated the power of our proposed association tests under various scenarios. We used significance level $\alpha = 2.5 \times 10^{-6}$, which corresponds to Bonferroni adjustment for testing of 20,000 genes. We first simulated samples of 5,000 individuals distributed in 3-generation pedigrees with 10 individuals each (Pedigree10 in **Figure 2.1**). Variants with frequency <1% (<5% where noted) explained 1% of the variance in a simulated quantitative trait. When all associated variants had the same effect size and the proportion of causal variants was small (~20%), SKAT had the largest power. When this proportion grew larger (~80%,

although this might not be very realistic, we increased the proportion to 80% for scientific investigation), VT became the most powerful test (**Table 2.1**). Although we did not simulate a relationship between frequency and effect size among causal variants, VT provided greater power because it sometimes excluded relatively common unassociated variants from consideration, reducing noise. When fraction of causal variants is small, methods that explicitly allow for heterogeneity in effect sizes do better, since no correlation between causality and effect size was simulated, VT can't easily exclude most of the unassociated variants. In practice, the true list of causal variants is usually unknown; and allele frequency is often a good proxy to identify variants likely to modify gene function [Nelson et al. 2012]. In a simplified scenario where only causal variants were grouped and other variants were discarded, the basic burden test became optimal (**Table 2.1**).

We next considered more complex scenarios. When 20% causal variants decreased trait values and the remainder increased trait values, the power of burden and VT tests dropped dramatically and SKAT became the most powerful test, regardless of the proportion of causal variants (**Table 2.2**). When we set up our simulation so that each variant explained the same fraction of trait variance (and, thus, so that rarer variants had larger effects), SKAT remained the most powerful test when the proportion of causal variants was small, but the Madson-Browning weighted burden (MB) test outperformed VT and SKAT when the proportion of causal variants was large (80%) (**Table 2.3**). This was expected since, in this setting, relative effect sizes match those predicted by the Madson-Browning weighting scheme.

Power when Misspecifying Frequency Threshold

We next investigated the impact of misspecifying frequency thresholds during analysis. **Figure S2.2A** shows that when causal variants have the same effect sizes, VT and Madson-Browning-weighted burden tests perform well as long as the frequency cut-off used during analysis is larger than the cutoff used for simulation. In contrast, the power of SKAT and simple burden tests is greatly reduced when incorrect frequency thresholds are used for analysis. **Figure S2.2B** shows that when rare causal variants have larger effects and all variants explain the same amount of trait variance, all tests reach maximum power at a frequency threshold less than or equal to 0.01, the threshold for simulating causal variants. Whereas the power of VT and MB remain close to optimal, the power of SKAT and the simple burden tests drops greatly as the frequency threshold used for analysis increases and non-causal and small effect variants enter the analysis. In real data analysis, because true disease model is unclear, we recommend multiple frequency thresholds should be used when using SKAT or simple burden tests [Lange et al. 2014].

Relative Power of Family Samples and Unrelated Individuals

We used simulations to compare the benefits of samples of families and unrelated individuals in association studies. Family samples can allow many copies of the same trait associated rare alleles to be observed in a single study. Variability in allele counts is larger in families, particularly in pedigrees with many descendants for each founder. For example, for a variant with allele frequency 0.0005 (~5 alleles expected when 5000 individuals are sequenced), the standard deviation of the allele

counts in a sample matching Pedigree50 (from **Figure 2.1**) is >3 times larger than a sample of unrelated individuals (see **Table S2.2** for details) – meaning that the chance of observing >10 copies of the variant is 20% when families matching Pedigree50 are sampled, but 4% in samples of unrelated individuals.

We speculated that the increased variability in allele counts in family samples would mean that family samples might sometimes hit a "jackpot" and sample many copies of a trait associated rare allele, increasing power. This speculation was supported by our simulations: a sample of 5,000 individuals in families matching Pedigree50 provides >2-fold greater power to detect a variant with frequency 0.001 and effect size 1 than a population sample of the same size (power was 0.9% in sample of unrelated individuals and 2.3% in sample of families, **Figure S2.3)**. This increase may seem paltry, but it is important to remember that many susceptibility loci underlie each human complex trait: if there are hundreds of such loci and power increases from 0.9% to 2.3% at each of those, the odds of a successful discovery will increase dramatically. The idea of "jackpot" effect was also supported by close examination of our simulation results. Among all 10,000 simulated samples, the average frequency of trait associated alleles was 0.0010, but in samples that have association p-value $<1x10^{-8}$, the frequency of trait associated alleles was higher, averaging 0.0032, a >3 fold increase. The relative advantages of family samples over unrelated samples decrease in settings where power (and, typically, the number of expected rare allele carriers) is high. For example, when sample size increases, allele

frequency increases, or effect size (or variance explained) increases unrelated samples quickly become more powerful (**Figure S2.3**).

Consistent with patterns in single variant association power, **Figure 2.2** shows that family studies have the similar advantages in studies of gene-level rare variant associations. For example, in a sample of 5,000 individuals, power to detect a gene where 20% of variants with frequency <1% are causal and explain 0.5% trait variance increases from 1% for unrelated individuals to 13% for family samples.

Advantages in power from studies of families are strongly correlated to the variance of allele counts (which is a function of family size and pedigree structure). For example, a sample of families matching Pedigree50 (**Figure 2.1**) has largest variance in allele counts (**Table S2.2**) and also the largest power for detecting a gene explaining 0.5% of trait variance in a sample of 5,000 individuals (**Figure 2.3**), whereas a sample of families matching Nuclear4 (**Figure 2.1**) has the smallest variance in allele counts and provides the smallest increase in power relative to samples of unrelated individuals (in this simulation, 20% of variants with frequency 1% were causal). All family samples have larger variance in allele counts than unrelated samples. For example, for a variant with frequency .1%, the standard deviation of the allele count in family samples with 5,000 individuals in Pedigree50 structure is 10.3, which is ~3 times the standard deviation of allele counts from unrelated samples of the same size.

The advantage of family samples extends to extremely rare variants. **Figure 2.4A** shows that when 20% of singleton variants (defined as alleles present only once in our initial pool of 10,000 simulated sequences) in a gene were causal explaining 0.5% trait variance, power to detect gene-level association increased dramatically from 3.5% in a study of 5,000 *unrelated* individuals to as much as 19.3% in a study of 5,000 *related* individuals. **Figure 2.4B** shows that when sample size increase to 10,000 individuals, the window where family samples are more advantageous becomes narrower.

In all examples highlighted so far, family studies outperform studies of unrelated individuals but in all of these examples power was low for both families and unrelated individuals. We expect that this is actually a common situation in human genetic studies – there may be very large numbers of trait associated loci but any single study may only provide enough power to detect a few of these. To explore this situation directly, we estimated power to detect at least one of several disease-associated loci. Assuming power to detecting association at a specific gene is p and x genes with similar effect variants exist, then the power to detecting at least one of these is $1-(1-p)^x$, assuming independent genes. **Figure 2.5** shows dramatic advantages in the power to detecting at least one of 20 trait associated genes, each explaining the same proportion of trait variance. For example, power to detect at least one gene explaining 0.5% trait variance (when 20% variants in the gene and with frequency <1% are causal) when 20 such genes exist is >90% in sample of 5,000 individuals distributed in families matching Pedigree50 (**Figure 2.1**), whereas

only ~20% in a sample of 5,000 unrelated individuals. The power advantage in family samples increases with the variability in allele counts, which in turn is driven by pedigree structure (**Figure 2.6**).

Families matching Pedigree50 are not easy to find. For a more realistic comparison of the power of studies of families and unrelated individuals, we repeated our simulations using the family structures and phenotypes observed in the SardiNIA sample. To preserve the correlation of phenotypes among family members, we started with observed HDL values together with sex, age and age-squared as covariates. **Figure S2.4** shows that the SardiNIA families provide larger power for discovering rare variants with moderate effect sizes than studies of same numbers of unrelated individuals. For example, the SardiNIA sample provides 1.6% power to detect a variant with frequency 0.0001 and effect of 2.5 trait standard deviation units, whereas unrelated samples provide only 0.05% power (**Figure S2.4A)**. If 100 such variants exist, the SardiNIA sample provides ~80% power to detect at least one, but an equal number of unrelated individuals provides only ~5% power to detect at least one of such a variant (**Figure S2.4B**). When allele frequency increases (**Figure S2.4C, S2.4D, S2.4E, S2.4F**), the SardiNIA sample is still advantageous when effect sizes are moderate.

Real Data Analysis Using SardiNIA and HUNT Studies

To evaluate our approach further, we meta-analyzed blood HDL levels for *11,556* individuals from the HUNT and SardiNIA studies (See **Table S2.3** for descriptive

statistics for traits). Overall, 93,831 and 76,828 sites were polymorphic in the HUNT and SardiNIA studies respectively, resulting in 117,958 polymorphic variants when combining the two studies **(Table S2.4**). Among those, 52,700 variants were shared in both studies (**Table S2.5**), 41,130 variants are unique to the HUNT study, and 24,128 variants are unique to the SardiNIA study (**Table S2.6**). Using our meta-analysis method, both shared and non-shared variants contribute to association signals.

We first generated summary statistics for each study adjusting for relatedness using empirical kinship matrices estimated from genotype data. Within each sample, test statistics were well calibrated with Genomic Control 1.00 in HUNT study and 1.01 in SardiNIA sample (See **Figure S2.5** for QQ plots). To illustrate the importance of taking into account phenotype correlations, consider that analyzing the SardiNIA exome chip data and treating the samples as unrelated results in a genomic control value of 1.45, which is unacceptably high (results not shown); but using our approach, genomic control becomes 1.01. We next proceeded to meta-analyze single variants. **Figure S2.5** shows that our meta-analysis statistics were also well calibrated with genomic control value <1.05, both for common and rare variants. At a significance threshold of $p<4.23 \times 10^{-7}$ (corresponding to 0.05/117,958*),* we found significantly associated low-frequency and rare variants at *CETP, LIPC, LIPG,* and *LPL* for HDL (MAF < 5%; See **Figure S2.6** for Manhattan plots). Significant rare variants were only found in *LIPC* and *LIPG* (MAF < 1%).

We then proceeded to gene-level meta-analyses. Again, test statistics appear well calibrated, with genomic control value <1.05 (See **Figure S2.7** for QQ plots). Also, by examining QQ plots from SardiNIA and HUNT study (See **Figure S2.7**), we discovered that, for family samples or samples from isolated population, in the analysis of rare variants, a small number of individuals can be quite influential such that all variants that are shared between this set of individuals (or families) will exhibit similar and often small p-values. This can lead to apparent inflation in QQ-plots, where confidence intervals are calculated assuming all statistics are independent. At a significance threshold of $p<2.84x10^{-6}$ (corresponding to 0.05/17,574 and thus allowing for the number of genes tested), we found association at *APOC3, CETP, LIPC, LIPG,* and *LPL* for HDL (See **Table 2.4** for tabulated results and **Figure S2.8** for Manhattan plots). Among those, *APOC3, LIPG,* and *LPL* had evidence of association stronger than the most significant single variant in the region. In *APOC3*, none of the individual low frequency and rare variants had p-value lower than 10^{-4} on its own (**Table 2.4**).

Comparison with other Methods and Tools

To validate our approach, we compared our implementation to several others in a simulated family sample of 10,000 individuals distributed across 1000 families matching Pedigree10 (see **Figure 2.1**). 4000 genes with 1,000 base-pair were simulated in families from a pool of haplotypes. A quantitative trait was simulated under the null. Variants with MAF<0.05 were grouped for gene-level tests. Pedigree-based kinship matrices were used in all analyses. We then analyzed the simulated

sample using our own famrvtest (SKAT, burden and VT tests), pedgene (burden and Kernel test) [Schaid et al. 2013], famSKAT [Chen et al. 2013], and FFBSKAT [Svishcheva et al. 2014]. **Figure S2.9** shows that all tests generate well-controlled QQ plots under the null.

To compare methods under the alternative, we simulated a dataset of 5,000 individuals (500 x Pedigree10) with a 1,000 base-pair long gene where 50% variants with MAF<0.05 were causal and together explained 1% trait variance. We simulated data sets where all causal variants had the same direction and also where half of the causal variants had opposite effects. In this simulation, our method always matched or slightly outperformed alternative implementations (see **Figure S2.10**).

These comparisons also allowed us to evaluate computation performance and requirements for our tool. Wherever possible, we tried to provide faster computation, less memory use, while still allowing for flexible input formats and varied choices of association tests. famrvtest is a command line tool implemented in C++. It uses computationally efficient algorithms to fit linear mixed models [Lippert et al. 2011], and recognizes pedigree-based kinship estimates as block-diagonal matrices to save computational effort. For our simulated dataset with 10,000 individuals and 164,323Z variants distributed across 4,000 genes, analysis with famrvtest required 1.5 hours and 1.3GB of memory to calculate both SKAT and

burden test statistics, a savings of up to 10-100 fold relative to alternative tools (see **Table S2.7**).

Discussion

Gene-level association tests and meta-analysis are important tools for discovering rare variant associations. We have proposed a series of methods that facilitate these analyses in family samples (or in samples where cryptic relatedness is modeled using variance components). Our C++ tools implement simple burden tests, weighted or un-weighted; and variable threshold tests as well as SKAT tests that outperform other tests when only small fractions of variants in each gene are causal or when variants with opposite effects reside in the same gene.

We compare the relative benefits of family samples and population samples. By simulation, we show that family samples can provide substantially greater power for rare variant association studies because of a "jackpot" effect – the potential for observing many copies of a trait associated rare variant. This advantage is likely to be extremely important in the first generation of rare variant association studies, each of which is only expected to detect a small fraction of all the true rare variant association signals. An example of successful discovery of such variant is rs72658864/V578A in *LDLR*, a rare variant associated to LDL with effect size 23.7 mg/dl [Sanna et al. 2011]. This variant was observed with frequency 0.00035 in the SardiNIA sample, where it was present in multiple families, but has not yet been observed in the 1000 Genomes [Abecasis et al. 2012] or the NHLBI Exome

Sequencing Projects [Fu et al. 2013; Tennessen et al. 2012] suggesting that it is rare indeed.

We demonstrate the utility of our methods by analyzing two samples with complex inter-relatedness. Meta-analysis of SardiNIA and HUNT resulted in a well-calibrated genomic control value of 1.02 and increased signal at many loci known to be associated with HDL – demonstrating the feasibility of including family samples in rare variant meta-analysis. We expect that meta-analysis will be useful not only for combining data across studies but also to facilitate analysis of large samples genotyped or sequenced across multiple platforms or analyzed using a single platform but in a batched manner.

We foresee several potential areas for refinement of our methods. For example, a limitation for our current approach to meta-analysis is that cross study relatedness and sample overlap are not modeled. In genome-wide studies, it may be possible to overcome this limitation by using the genome-wide correlation of test statistics between pairs of studies to calculate an adjustment factor that could account for overlap or relatedness between individuals in two studies [Lin and Sullivan 2009]– as suggested by Lin et al. for single marker meta-analyses. Extension of this idea has also been proposed in Han et al. 2013. Extending our methods to non-coding variants will also be attractive, particularly since the majority of trait-associated variants found to date are located in non-coding regions. A difficulty will be the development of good grouping strategies for non-coding variants, where

interpretation of functional consequence is more challenging. Another challenge we foresee is the extension of our methods to discrete traits. The natural way to do this is to consider an underlying continuous liability scale and use multivariate integration to fit the model, but there may be more computationally efficient alternatives to be discovered.

In summary, we have proposed a series of gene-level association tests for family samples and methods for calculating these in a meta-analysis of related and/or unrelated samples. We also implemented our methods in freely available and open source C++ tools: http://genome.sph.umich.edu/wiki/FamRvTest and http://genome.sph.umich.edu/wiki/RAREMETAL. We hope these tools and methods will facilitate the next round of gene-mapping studies.

Tables

Table 2. 1 Power when Causal Variants All Increase Trait Values and Have the Same Effect Sizes

MAF Cutoff	Causal Percentage	Group by MAF Cutoff				Group Only Causal Variants[b]			
		Burden	**Madsen-Browning**	**VT**	SKAT[a]	**Burden**	**Madsen-Browning**	**VT**	**SKAT**
0.01	20%	9.7	3	13.1	**36.6**	**94.3**	86.7	92.9	82.6
	80%	82.4	64.7	**88.1**	61	**96**	82.1	94.3	70.7
0.05	20%	14.6	2.6	24.9	**36.3**	**95.4**	75.3	93.8	86.5
	80%	81.3	39.5	**89.2**	75	**96.3**	55.3	94.3	82.9

Simulated samples each had 5,000 individuals, organized in families with pedigree10 structure (See **Figure 1**). Causal variants were selected among those identified in simulated 1,000 base-pair sequences and explained 1% of trait variance. Each causal variant had the same effect size and direction. Power is tabulated as a percentage of simulations exceeding significance threshold. Significance level $\alpha = 2.5 \times 10^{-6}$ was used in all simulations.

a. Power calculated from Madsen-Browning weighted SKAT.
b. Power when grouping only causal variants. This column represents the largest power we can achieve for each simulation setting.

Table 2. 2 Power Comparison when Causal Variants Can Have Opposite Effects

MAF Cutoff	Causal Percentage	Group by MAF Cutoff				Group Only Causal Variants			
		Burden	Madsen-Browning	VT	SKAT	Burden	Madsen-Browning	VT	SKAT
0.01	20%	4.6	0.4	6.0	**36.7**	38.9	21.1	43.4	**83.2**
	80%	30.5	10.4	33.4	**60.0**	42.6	18.8	42.2	**69.0**
0.05	20%	11.7	1.3	15.0	**35.7**	55.4	22.3	58.3	**88.3**
	80%	44.0	7.8	47.1	**74.7**	55.1	12.2	54.3	**81.6**

Simulated samples each had 5,000 individuals, organized in families with pedigree10 structure (See **Figure 1**). Causal variants were selected among those identified in simulated 1,000 base-pair sequences and explained 1% of trait variance. Among causal variants, 20% were randomly selected to be trait-decreasing, and the rest causal variants were trait-increasing. Power is tabulated as a percentage of simulations exceeding significance threshold. Significance level $\alpha = 2.5 \times 10^{-6}$ was used in all simulations.

Table 2. 3 Power Comparison when Causal Variants All Increase Trait Values and Explain the Same Amount of Trait Variance

MAF Cutoff	Causal Percentage	Group by MAF Cutoff				Group Only Causal Variants			
		Burden	Madsen-Browning	VT	SKAT	Burden	Madsen-Browning	VT	SKAT
0.01	20%	4.3	4.2	9.1	**20.8**	88.7	**94.9**	90.8	67.0
	80%	66.9	**86.6**	85.4	20.1	85.5	**97.1**	93.8	27.0
0.05	20%	3.8	5.1	9.3	**9.8**	78.8	**98.0**	90.1	53.0
	80%	38.6	**88.5**	82.1	9.4	56.0	**97.9**	92.6	12.4

Simulated samples each had 5,000 individuals, organized in families with pedigree10 structure (See **Figure 1**). Causal variants were selected among those identified in simulated 1,000 base-pair sequences and explained 1% of trait variance. Each causal variant explained the same amount of trait variance. All causal variants were trait-increasing. Power is tabulated as a percentage of simulations exceeding significance threshold. Significance level $\alpha = 2.5 \times 10^{-6}$ was used in all simulations.

Table 2. 4 Significant Genes from Gene-level Meta Analysis of HUNT and SardiNIA Exome Chip Data (HDL)

Gene	Burden	Madsen-Browning	VT (Actual MAF Cutoff)	SKAT[c]	Variants Included[d]	MAF	Effect Sizes (SD)	Single Variant p-values
APOC3[b]	2.3×10^{-6}	1.9×10^{-6}	6.4×10^{-6} (6.1×10^{-4})	4.5×10^{-5}	11:116701560:G:A	4.8×10^{-4}	0.959	1.4×10^{-3}
					11:116701353:C:T	5.6×10^{-4}	1.009	1.5×10^{-3}
					11:116701354:G:A	6.1×10^{-4}	0.528	5.7×10^{-2}
CETP	6×10^{-20}	2.7×10^{-3}	2.4×10^{-19} (3.2×10^{-2})	1.2×10^{-20}	16:57015091:G:C	3.2×10^{-2}	-0.359	1.3×10^{-20}
					16:57007387:C:T	4.3×10^{-5}	2.241	2.3×10^{-2}
					16:56995935:C:G	4.3×10^{-5}	-1.572	1.1×10^{-1}
					16:57012039:G:A	4.3×10^{-5}	-0.803	4.2×10^{-1}
					16:57009022:G:A	1.7×10^{-4}	0.309	5.3×10^{-1}
					16:57015076:G:A	2.2×10^{-4}	0.144	7.4×10^{-1}
					16:57012094:A:G	4.3×10^{-5}	0.182	8.5×10^{-1}
LIPG[b]	1.3×10^{-10}	6.7×10^{-9}	4.5×10^{-10} (9.4×10^{-3})	1.9×10^{-8}	18:47109955:A:G	9.4×10^{-3}	0.375	4.5×10^{-8}
					18:47113165:C:T	9.1×10^{-4}	0.668	2.3×10^{-3}
					18:47109939:G:A	1.7×10^{-4}	1.012	3.9×10^{-2}
					18:47101838:G:A	4.3×10^{-5}	1.000	3.1×10^{-1}
LPL[b]	3.7×10^{-11}	4.5×10^{-5}	1.2×10^{-10} (2.0×10^{-2})	2×10^{-11}	8:19813529:A:G	2.0×10^{-2}	-0.273	1.3×10^{-8}
					8:19805708:G:A	1.1×10^{-2}	-0.254	7.5×10^{-5}
					8:19816888:C:T	1.1×10^{-3}	0.234	2.3×10^{-1}
					8:19819628:T:G	4.3×10^{-5}	0.193	8.4×10^{-1}
LIPC	1.8×10^{-4}	1.5×10^{-4}	3.2×10^{-5} (6.2×10^{-3})	1.7×10^{-7}	15:58855748:C:T	6.2×10^{-3}	0.539	4.9×10^{-10}
					15:58837989:G:A	7.4×10^{-4}	0.542	2.5×10^{-2}
					15:58833993:G:A	3.1×10^{-2}	0.054	1.7×10^{-1}
					15:58830716:G:A	8.7×10^{-5}	0.123	8.6×10^{-1}
					15:58853079:A:C	5.9×10^{-3}	-0.003	9.8×10^{-1}
					15:58860956:G:A	4.3×10^{-5}	0.025	9.8×10^{-1}

Significance level 2.84×10^{-6} was used for reporting significant genes. Non-synonymous, splice, and stop variants with MAF<0.05 were included in analysis.

b: The gene-level p-value is smaller than the p-value for each of the single variants included in the test.

c: P-values of SKAT were generated using weights suggested in Wu el al. [Wu et al. 2011].

d: Variants are in the following format: CHR:POS:REF:ALT.

Figures

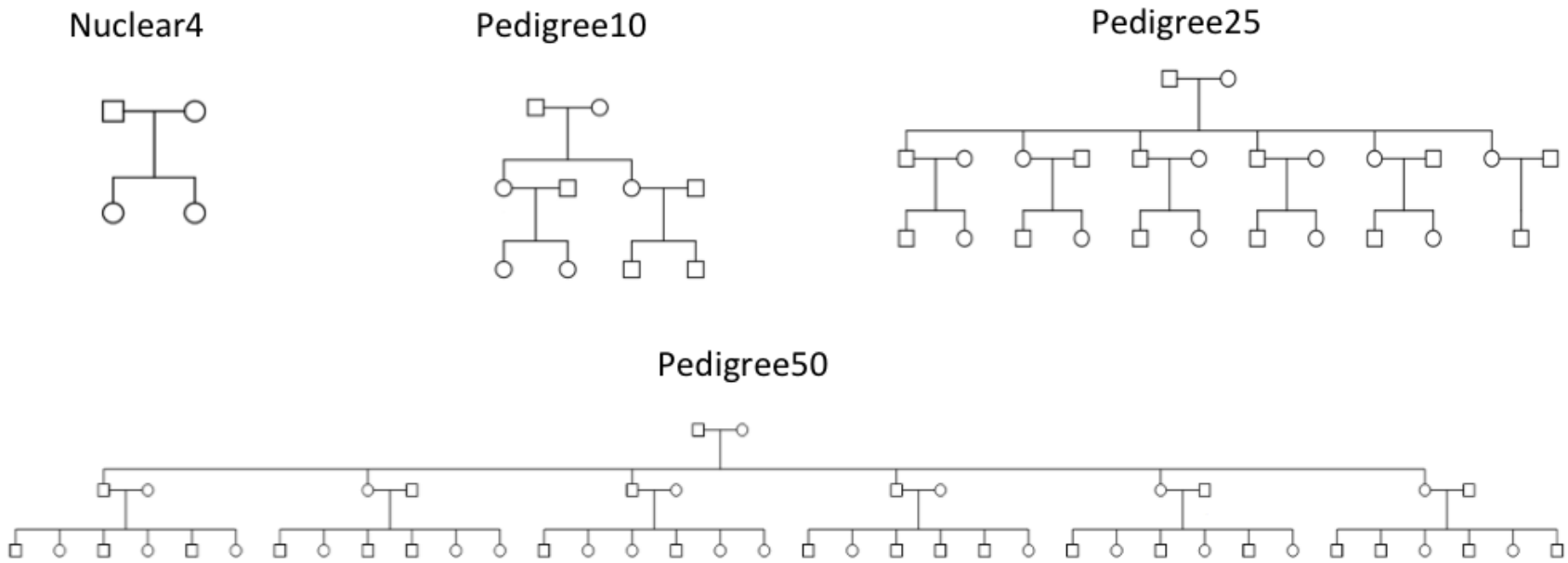


Figure 2. 1 Pedigree Structures Used in Simulations

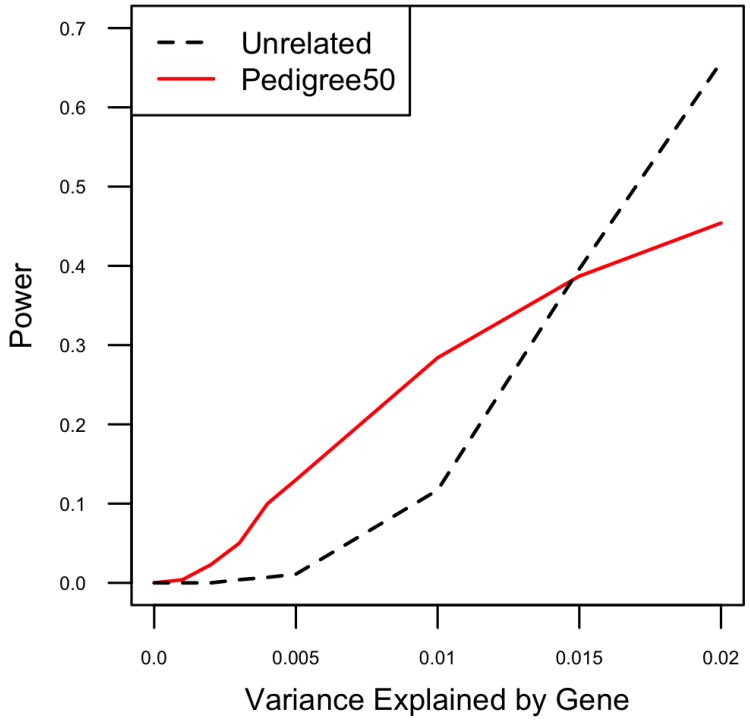


Figure 2. 2 Power to Detect Gene-level Association in Family and Population Samples

All samples had 5,000 individuals. All family samples used the Pedigree50 structure (see **Figure 1** for details). In every simulation, 10,000 haplotypes were simulated and 20% of variants with MAF<0.01 were randomly selected as causal variants, each explaining the same amount of trait variance. Then, a subset of simulated haplotypes were selected as founder haplotypes, segregated through families according to Mendel's laws, and used to simulate quantitative traits. Power of the SKAT test was evaluated using 10,000 simulations and significance level $\alpha = 2.5 \times 10^{-6}$.

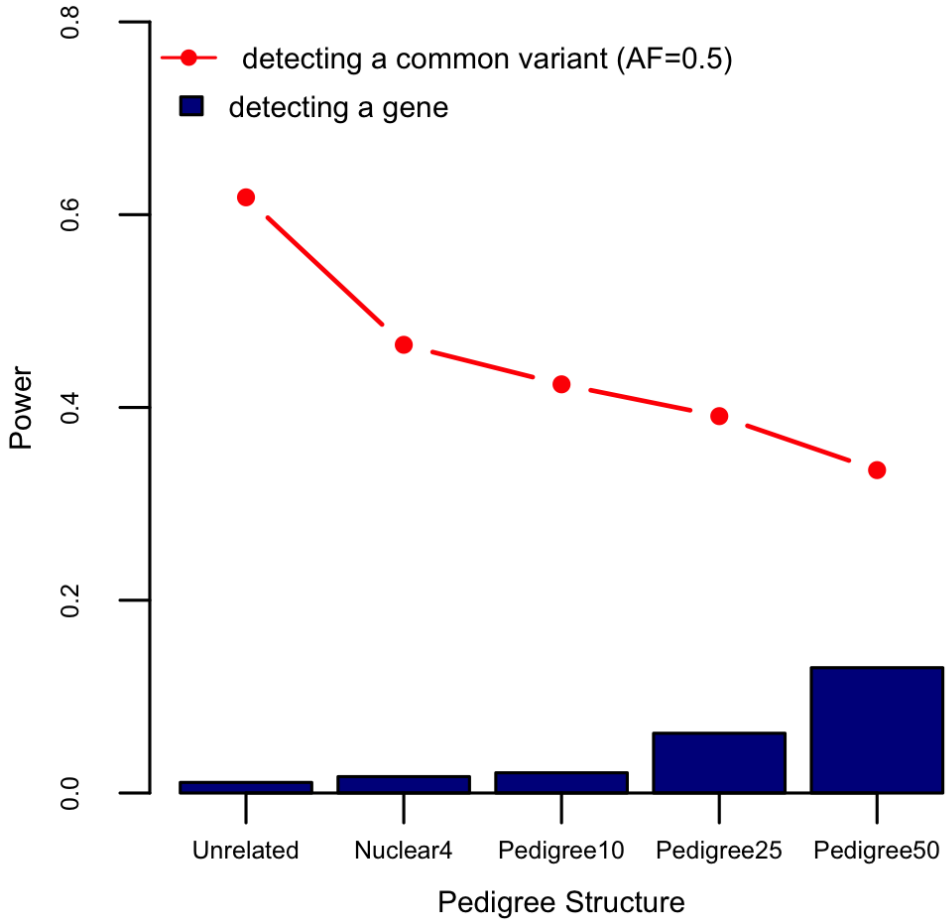


Figure 2. 3 Power to Detect Gene-level Association as a Function of Pedigree Structure

In each simulation, 20% of variants with MAF<0.01 were randomly assigned as causal, each explaining the same amount of trait variance. Together, causal variants explained 0.5% of trait variance. For comparison, the red line shows the power for one variant with frequency of 0.5 and explaining 0.5% of the trait variance. Power of the SKAT test was evaluated using 10,000 simulations and significance level $\alpha = 2.5 \times 10^{-6}$.

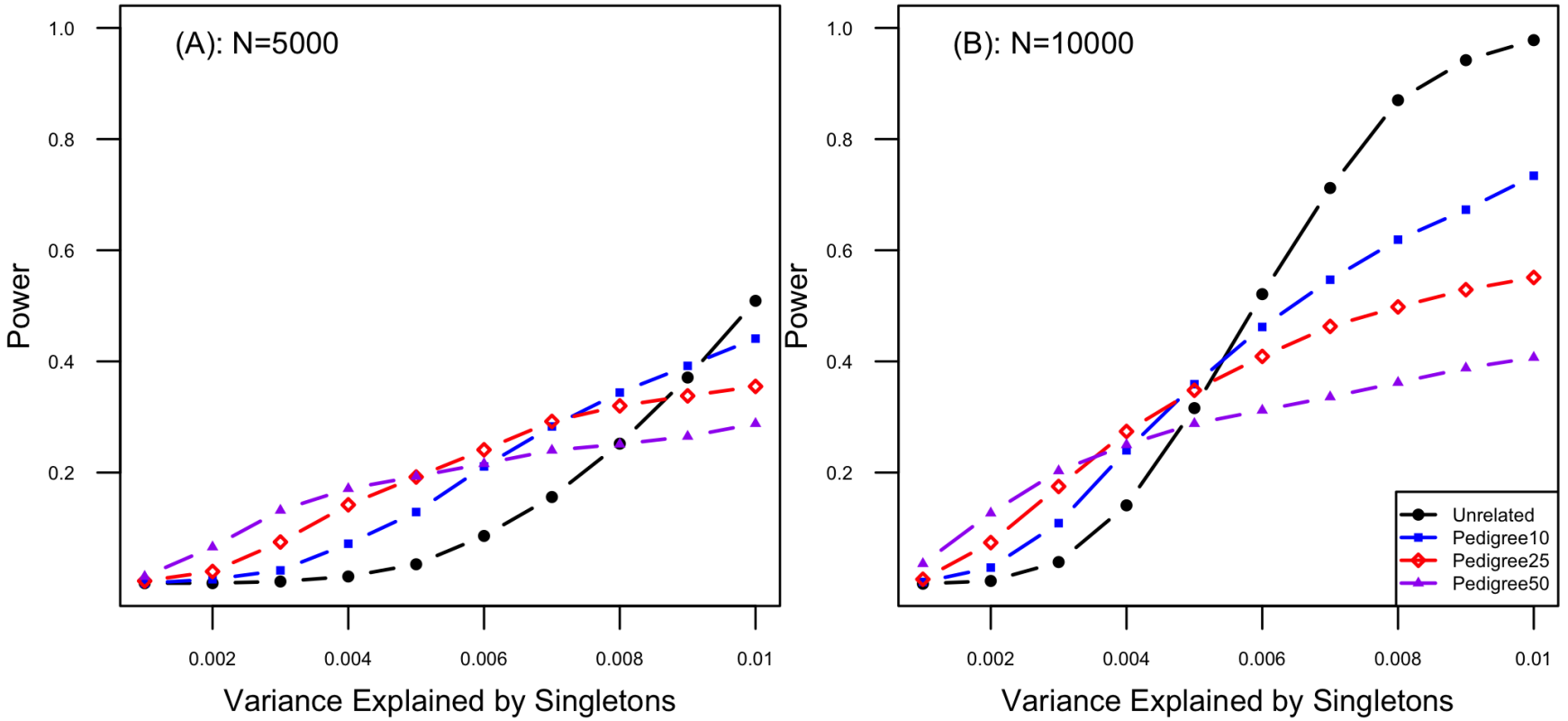


Figure 2. 4 Power to Detect Gene-level Association When Singletons are Causal

In each simulation, 10,000 simulated haplotypes were simulated. 20% singletons from these haplotypes were chosen as causal variants, together explaining various proportions of trait variance. Trait heritability was 40%. Then, a subset of haplotypes were used to seed founder haplotypes in each family sample. Only singletons or private variants were grouped for association tests. 10,000 simulations were used to evaluate power in samples of 5,000 individuals (panel A) or 10,000 individuals (panel B). See **Figure 1** for details of pedigree structures. Power was evaluated in 10,000 simulations using significance level $\alpha = 2.5 \times 10^{-6}$.

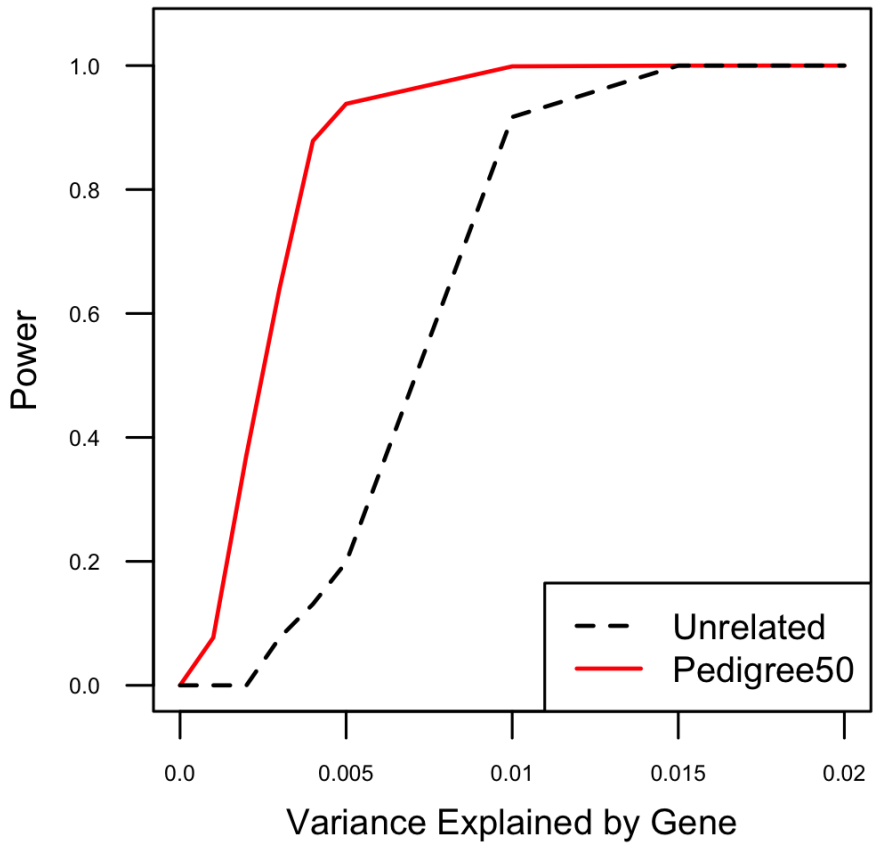


Figure 2. 5 Power to Detect at least one of Twenty Causal Genes

Assuming power to detect association at a specific gene is *p* and *n* genes with similar effect variants exist, then the power to detect at least one of these is $1-(1-p)^n$. See Figure 2 for power to detect a single gene and additional details of simulation settings.

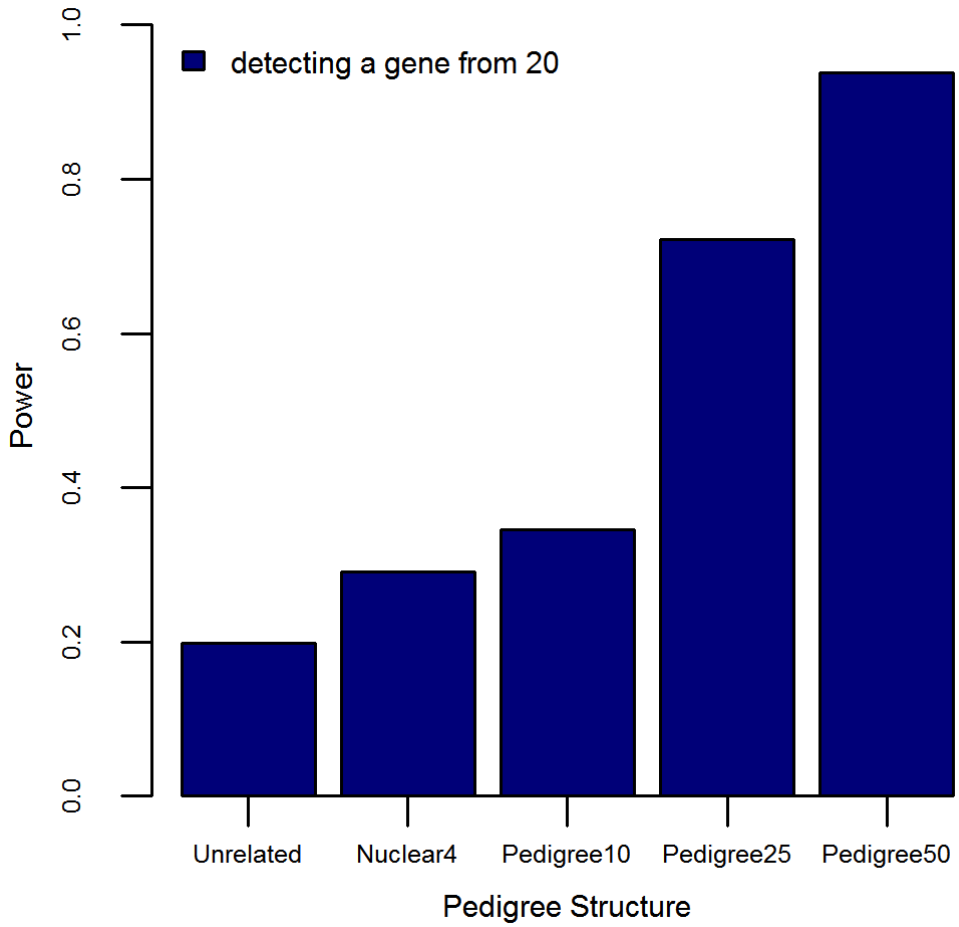


Figure 2. 6 Power to Detect at least One of Twenty Causal Genes as a Function of Pedigree Structure

The blue bars show power to detect at least one gene where rare variants explain 20% of trait variance and 20 such genes exist. The red line shows the power to detect at least one common variant with frequency 0.5 that explains 0.5% of trait variance when 20 such variants exist. See the legends of Figure 3 for simulation settings. See the legends of Figure 5 for calculating power to detect at least one of *n* genes with similar effect variants exist.

Supplementary

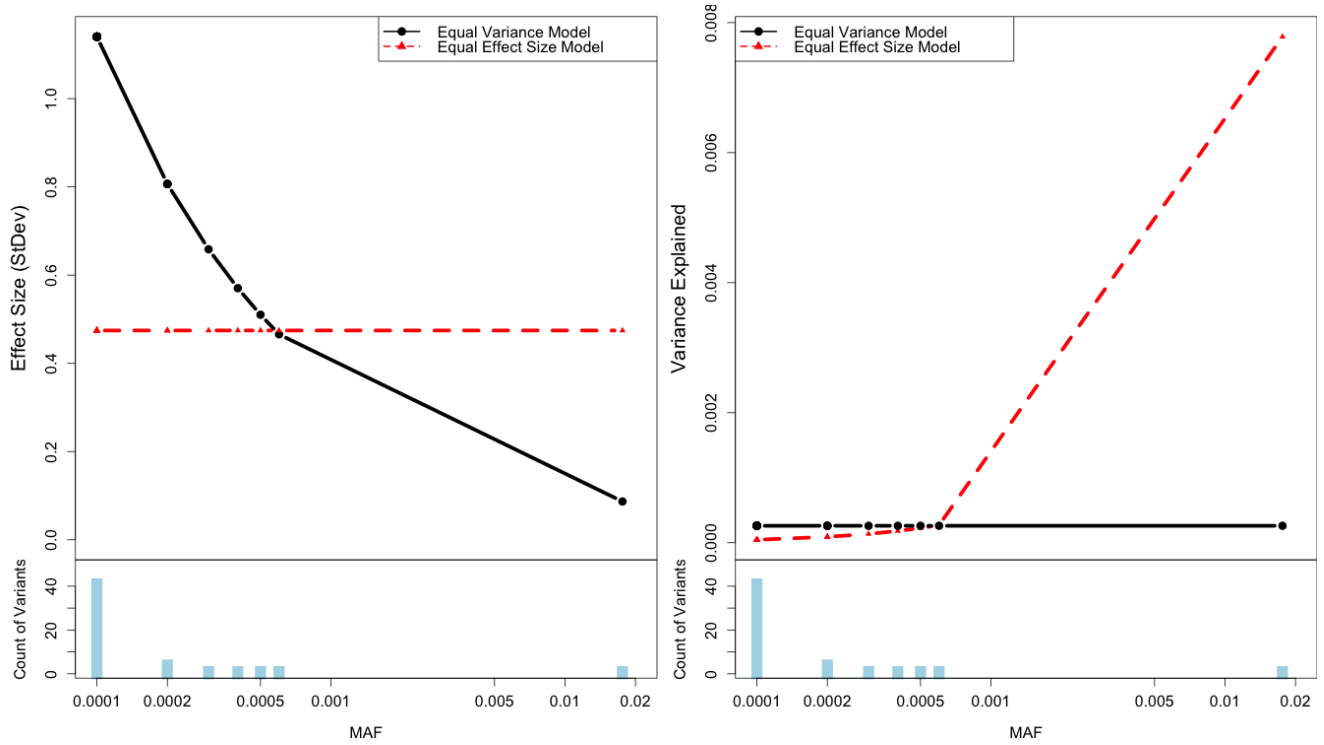


Figure S2. 1 Example Effect Sizes and Variable Explained in Equal Effect Size Model and Equal Variance Model

80% variants with MAF<0.05 from one simulated population sample of 5,000 individuals were randomly selected as causal. Bars of count of variants represent the actual count of causal variants with a certain minor allele frequency.

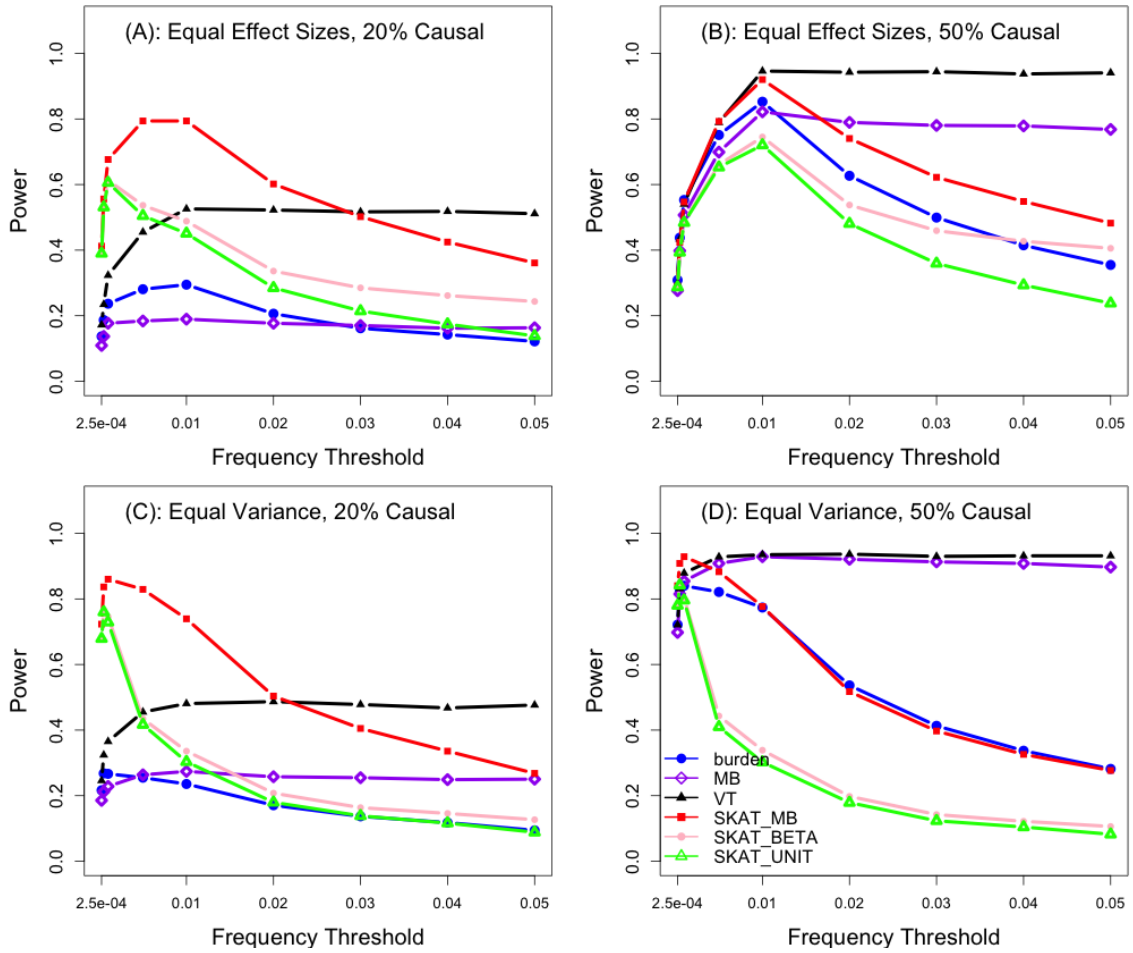


Figure S2. 2 Power of Detecting Association of a Gene when Grouping Rare Variants Using Various Frequency Thresholds

(A) and (B) Causal variants have the same effect sizes. (C) and (D) Causal variants explain the same amount of variance. All samples have pedigree10 structure with 5,000 individuals. 20% or 50% variants below frequency 1% were selected as causal variants that are unidirectional, explaining 2% of trait variance. Then various frequency thresholds were used to form the group of variants to test upon. Power was evaluated using 1000 simulations.

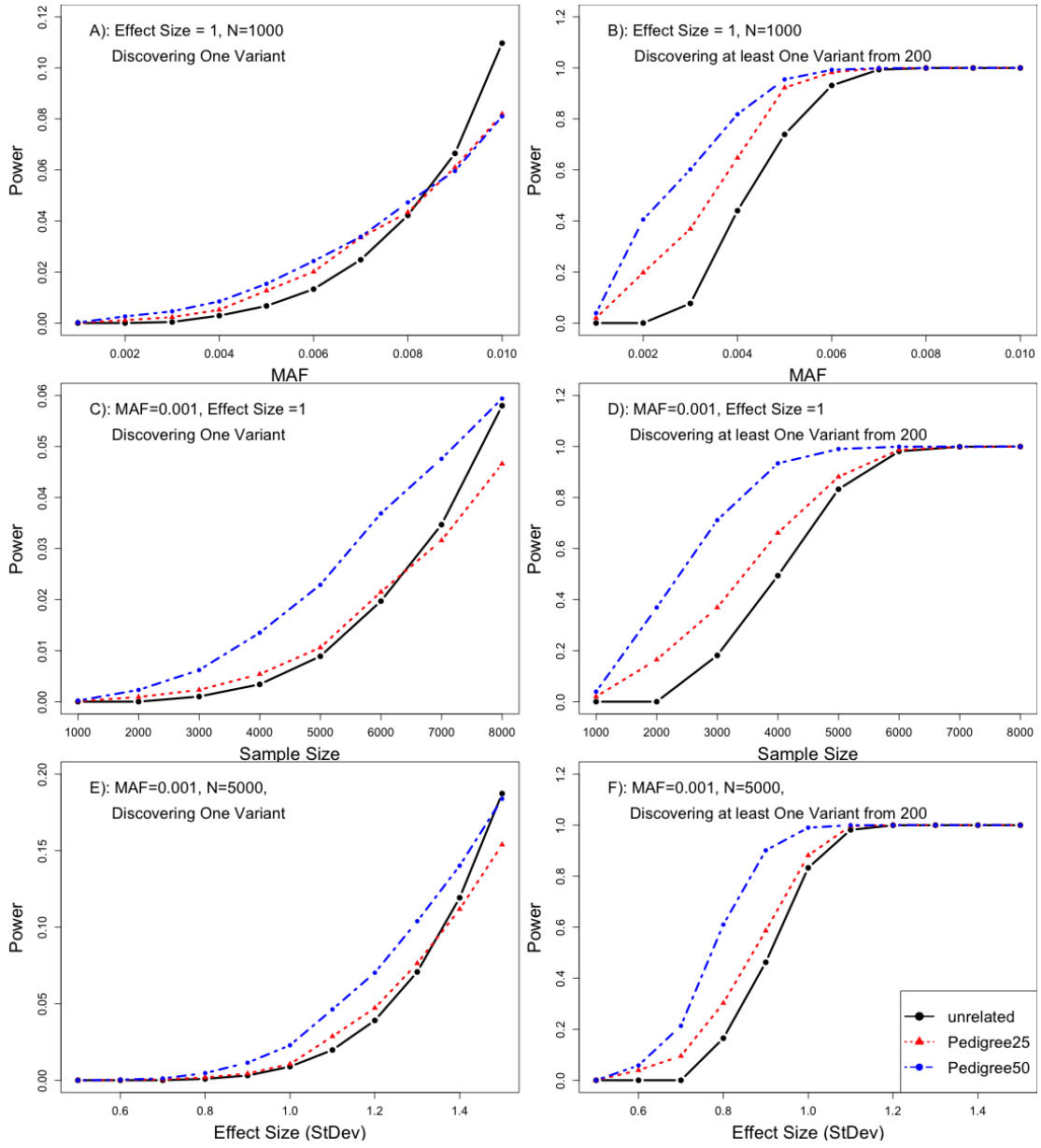


Figure S2. 3 Power of Rare Variant Associations in Family Samples and Population Samples

To simulate genotype, 16,000 founder haplotypes of a single variant with certain allele frequency were simulated and then gene-dropped to children in family samples (for samples with less than 8,000 founders, founder haplotypes were randomly selected from the pool of 16,000). Then genotypes of each individual were calculated as the count of rare alleles. Null phenotypes were simulated at first: in unrelated samples, null phenotypes were random draws from a normal distribution; in family samples, null phenotypes were simulated such that family members have correlated trait values (with heritability 0.4) based on covariance matrix. Then, final phenotypes were modified based on genotype and effect sizes: individuals with rare alleles have phenotypes added the amount of *effect size* ×*genotype*. A), C), and E) show the power of detecting a single variant. B), D), and F) show the power of discovering at least one variant assuming 200 trait-associated variants of the same frequency and effect size exist. We used $\alpha=1x10^{-8}$ for power.

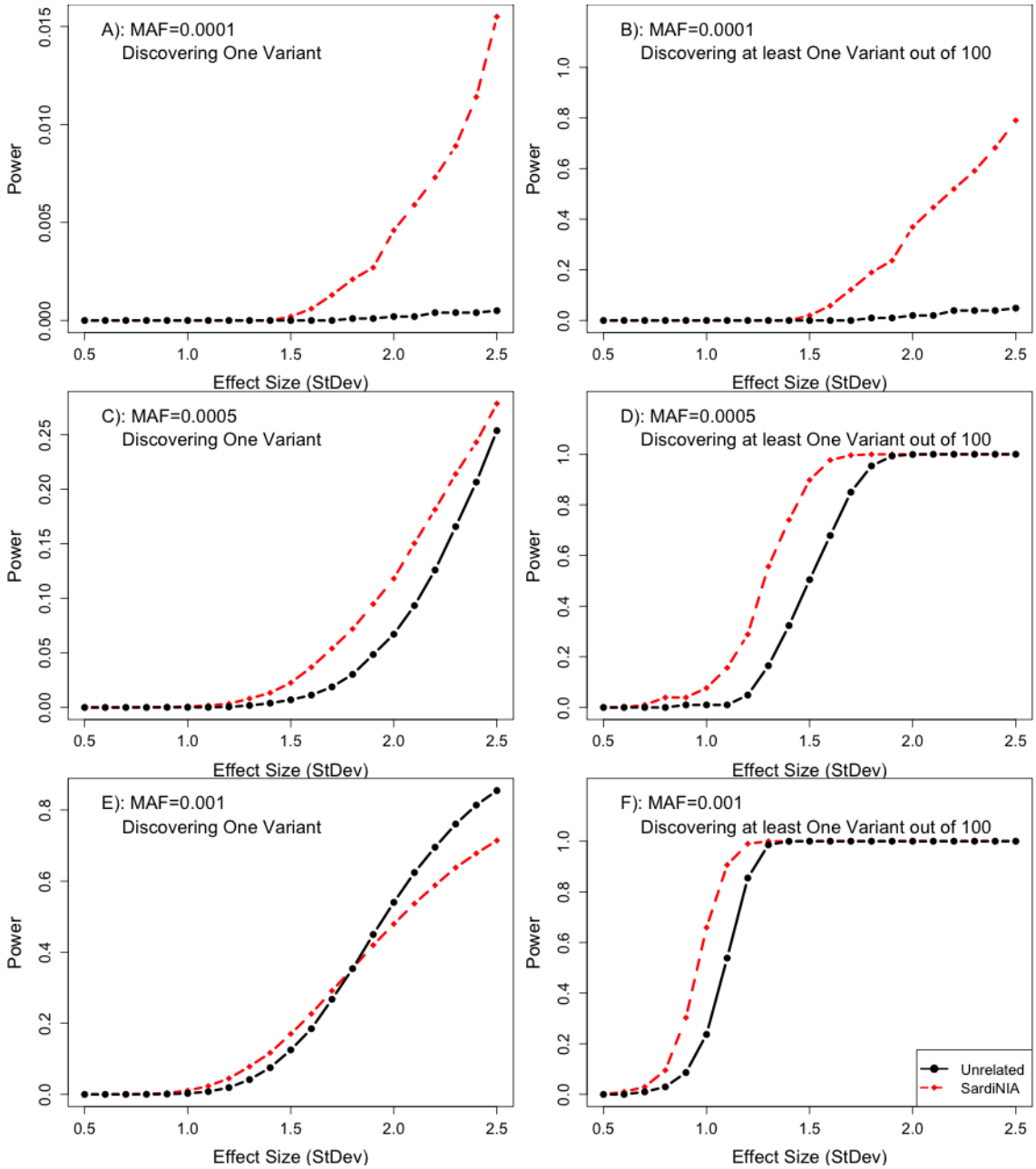


Figure S2. 4 Power of Discovering a Single Variant of Various Effect Sizes in SardiNIA Sample and Population Sample

SardiNIA sample has 5,916 individuals with both HDL and covariates (age, age2, and sex) measured. To compare power with SardiNIA sample, unrelated samples of size 5,916 were simulated. To simulate genotypes, 11,832 haplotypes of a single variant were simulated and assigned to individuals in the unrelated sample; then, a subset of 7,222 haplotypes were randomly selected as founder haplotypes and dropped to children in the SardiNIA sample by Mendelian inheritance. Phenotypes were simulated using the same method described in **Figure S2.3**. Power was calculated using 10,000 simulations. A), C) and E) show the power of discovering a single variant with frequency 0.0001. B), D), and F) show the power of discovering one variant assuming 100 variants of the same frequency and effect size exist. We used $\alpha=1x10^{-8}$ to determine power.

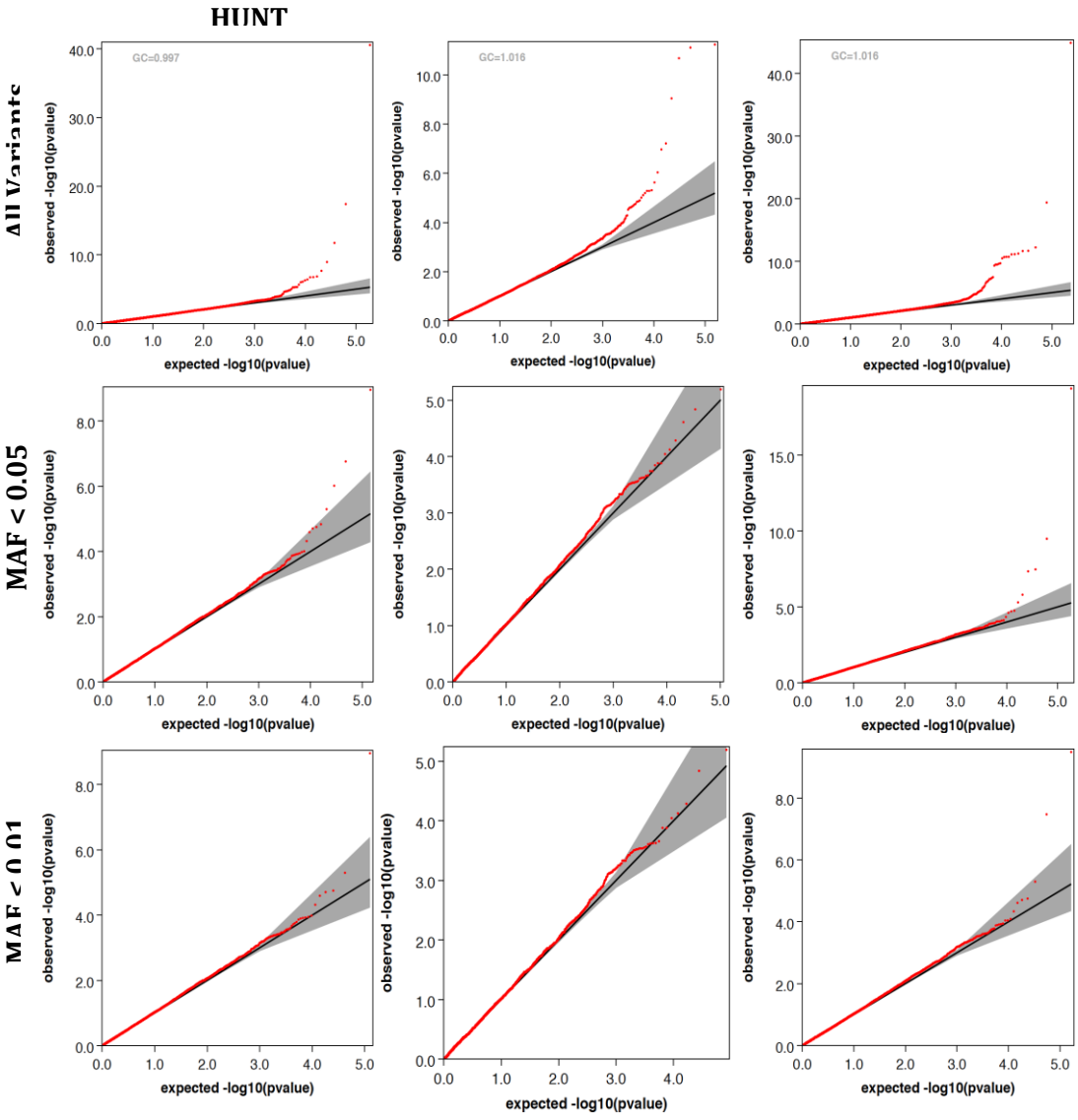


Figure S2. 5 QQ Plots of HUNT, SardiNIA, and Meta-analysis Single Variant Associations

Variants with HWE p-value < 1.0e-05 or call rate < 0.95 were excluded from the analyses.

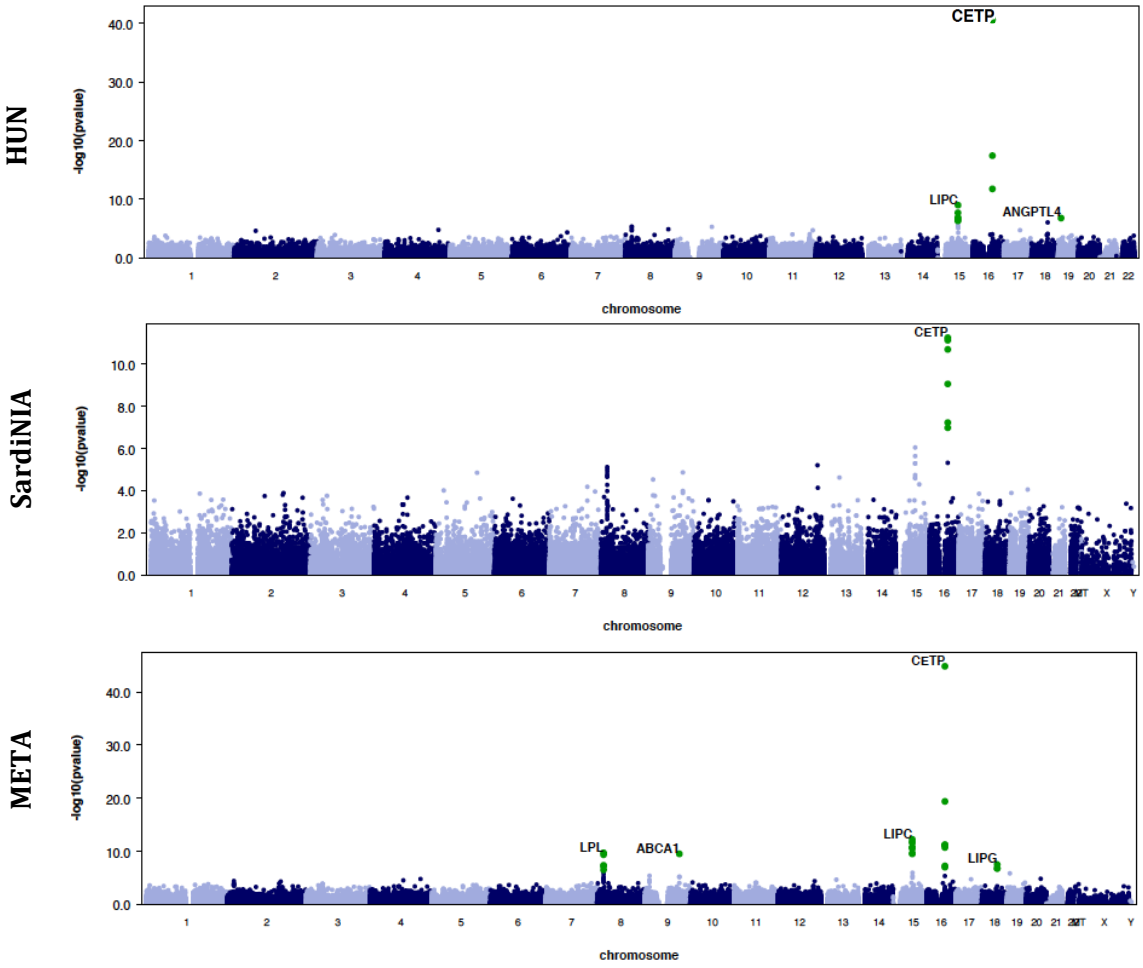


Figure S2. 6 Manhattan Plots of HUNT, SardiNIA, and Meta-analysis Single Variant Associations

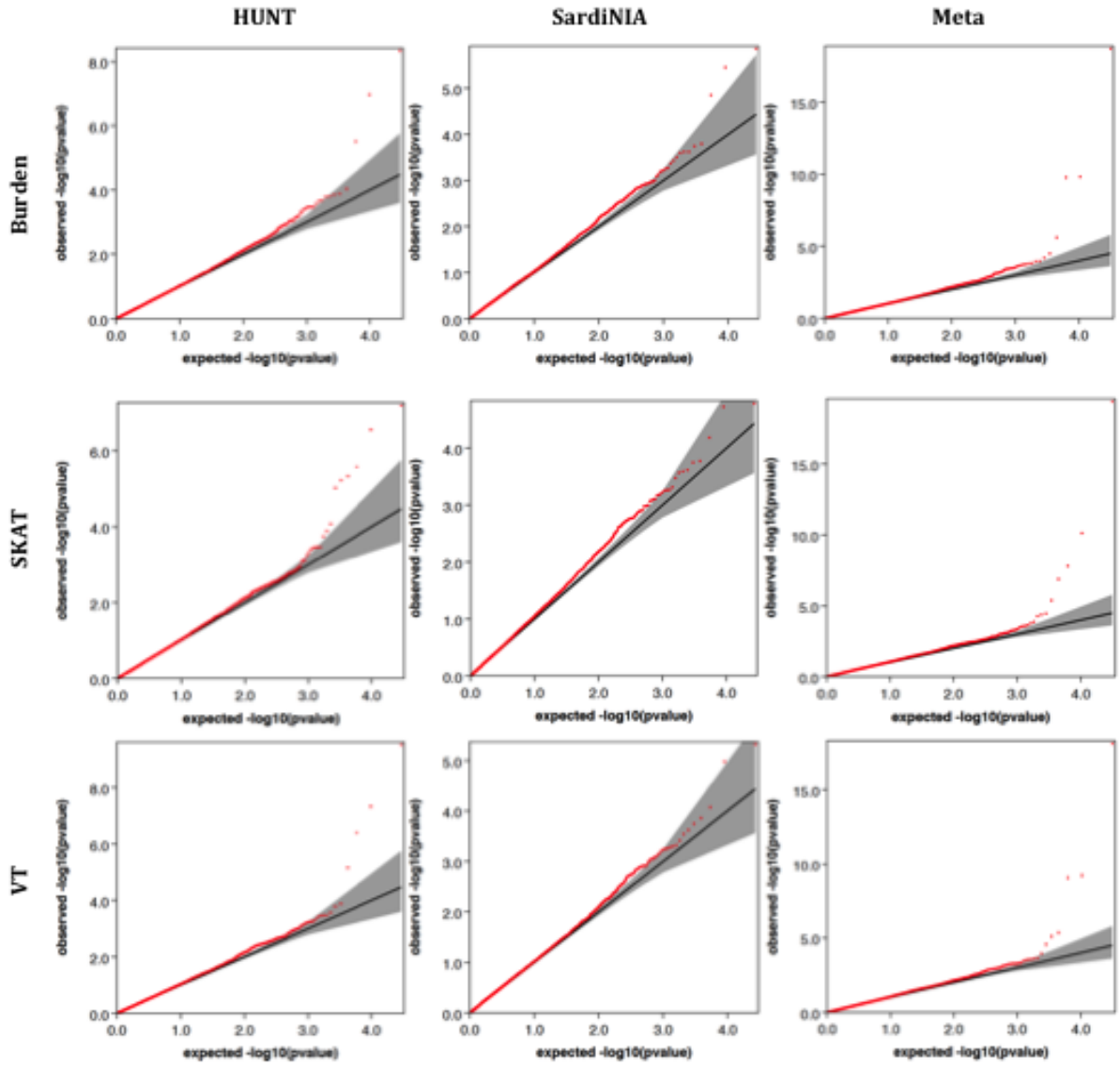


Figure S2. 7 HUNT SardiNIA and Meta-analysis Gene-level Associations QQ Plots

Variants with MAF<0.05 were grouped for gene-level tests. Variants with HWE p-value < 1.0e-05 or call rate < 0.95 were excluded.

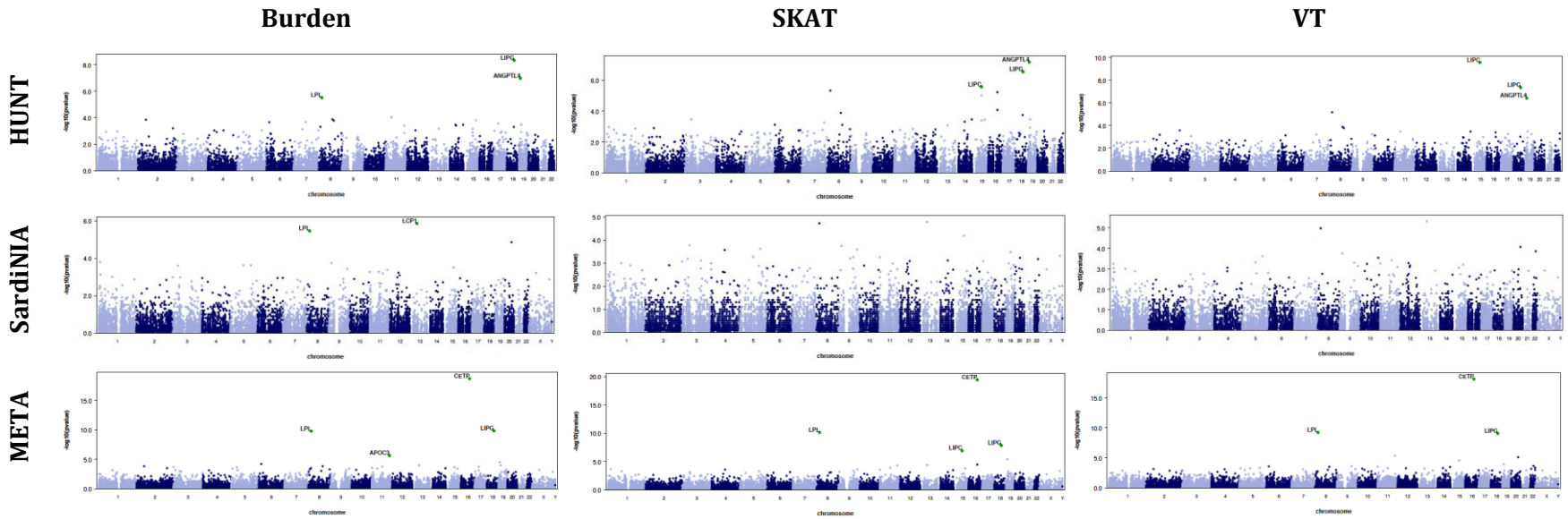


Figure S2. 8 Manhattan Plots for HUNT and SardiNIA Gene-level Associations

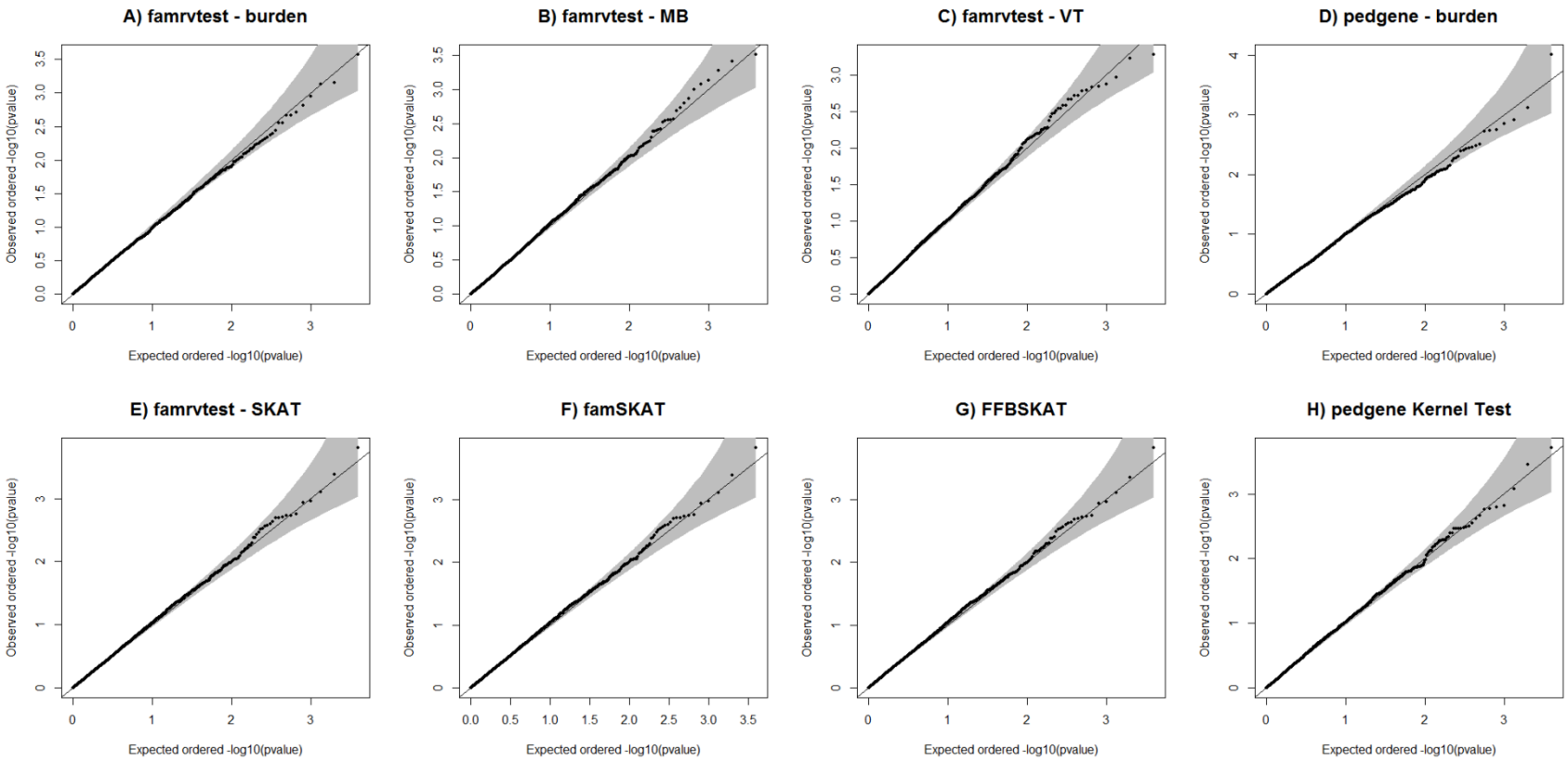


Figure S2. 9 QQ Plots Generated by Various Tools from Various Family-based Gene-level Tests Analyzing a Simulated Data Set

A), B), C), and D) are from VT and un-weighted and Madson-Browning-weighed burden and tests. E), F), G), and H) are from family-based SKAT and kernel tests. 4000 genes of 1k bp were simulated in a family sample with 10,000 individuals in 1000 families. A quantitative trait was simulated under the null. All analyses used pedigree-based kinship matrix. Davies method was used to calculate pvalue from SKAT test in famSKAT.

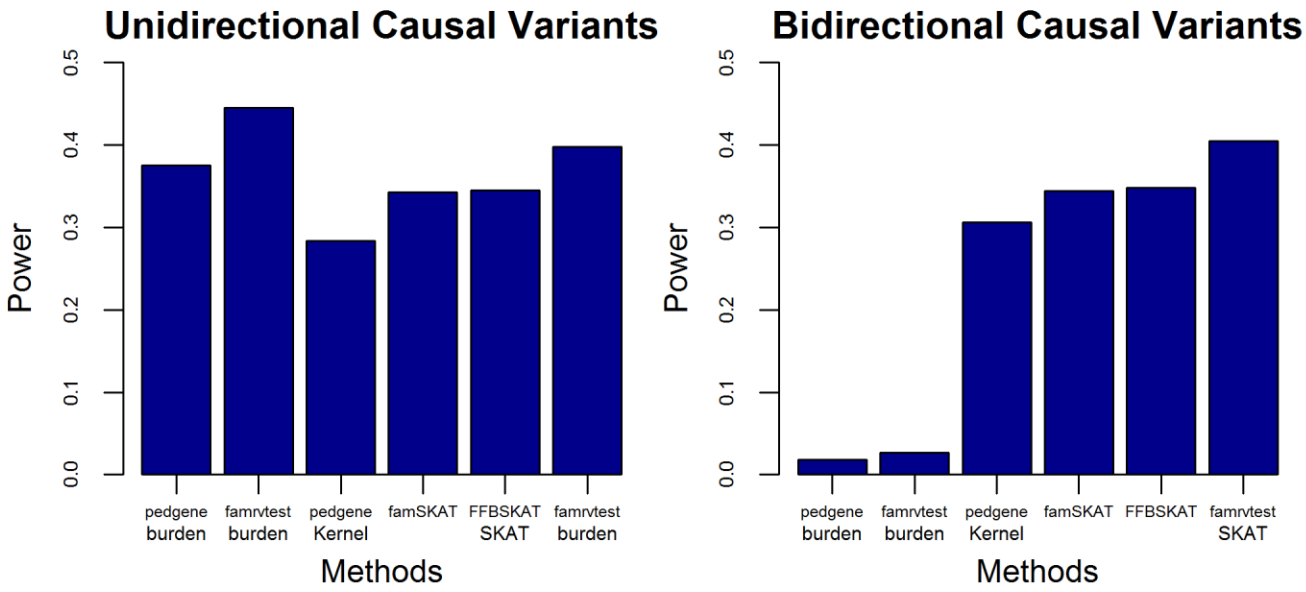


Figure S2. 10 Power Comparison among Various Methods and Implementations

1000 simulations of 5,000 individuals in 500 families were used to evaluate power for each method and implementation. A gene of 1000 base-pair explaining 1% trait variance was simulated in each data set. 50% variants with MAF<0.05 were selected as causal variants. In bidirectional scenario, half causal variants were selected to have opposite effects.

Table S2. 1 Type I Error of Gene-level Association Tests

a. MB is the Madsen-Browning weighted burden test.

Pedigree	N	Method	0.05	1×10^{-4}	1×10^{-5}	2.5×10^{-6}
95% Confidence Intervals[c]			(0.0498, 0.050)	($9.12\text{x}10^{-5}$, $1.09\text{x}10^{-4}$)	($7.23\text{x}10^{-6}$, $1.28\text{x}10^{-5}$	($1.11\text{x}10^{-6}$, $3.89\text{x}10^{-6}$)
Pedigree10	1000	Burden	0.050	9.06×10^{-5}	7.84×10^{-6}	1.67×10^{-6}
Pedigree10	1000	MB[a]	0.050	9.19×10^{-5}	6.84×10^{-6}	1.67×10^{-6}
Pedigree10	1000	VT	0.050	7.54×10^{-5}	7.34×10^{-6}	1.17×10^{-6}
Pedigree10	1000	SKAT[b]	0.050	7.99×10^{-5}	8.51×10^{-6}	2.50×10^{-6}
Pedigree10	5000	Burden	0.050	9.95×10^{-5}	8.00×10^{-6}	1.75×10^{-6}
Pedigree10	5000	MB	0.050	9.38×10^{-5}	8.25×10^{-6}	1.00×10^{-6}
Pedigree10	5000	VT	0.050	8.20×10^{-5}	8.25×10^{-6}	1.75×10^{-6}
Pedigree10	5000	SKAT	0.050	8.73×10^{-5}	9.00×10^{-6}	2.50×10^{-6}
Pedigree50	1000	Burden	0.049	8.43×10^{-5}	7.00×10^{-6}	1.33×10^{-6}
Pedigree50	1000	MB	0.049	8.02×10^{-5}	6.83×10^{-6}	1.50×10^{-6}
Pedigree50	1000	VT	0.050	8.03×10^{-5}	6.83×10^{-6}	1.83×10^{-6}
Pedigree50	1000	SKAT	0.048	8.28×10^{-5}	6.33×10^{-6}	1.33×10^{-6}
Pedigree50	5000	Burden	0.050	1.01×10^{-4}	1.02×10^{-5}	1.50×10^{-6}
Pedigree50	5000	MB	0.050	9.95×10^{-5}	9.00×10^{-6}	1.75×10^{-6}
Pedigree50	5000	VT	0.050	9.03×10^{-5}	8.75×10^{-6}	2.25×10^{-6}
Pedigree50	5000	SKAT	0.050	9.83×10^{-5}	8.50×10^{-6}	2.50×10^{-6}

b. SKAT uses the Beta(MAF,1,25) density weight.
c. 95% confidence interval calculated based on 5,000,000 simulations.
All estimates are within 95% confidence interval or below the lower
level of the 95% confidence interval.

Table S2. 2 Allele Counts in Population and Family Samples by Frequency

Sample	N (Founder)	Allele Counts (StDev)			
		MAF= **0.05**	**0.01**	**0.005**	**0.001**
Unrelated	**5000 (5000)**	500.2 (21.4)	100.2 (9.7)	50.0 (6.7)	10.0 (3.2)
Nuclear4	**5000 (2500)**	499.8 (32.1)	100.3 (14.6)	49.9 (10.7)	10.0 (4.8)
Pedigree10	**5000 (2000)**	499.2 (38.3)	100.7 (17.7)	50.1 (12.7)	9.8 (5.6)
Pedigree25	**5000 (1600)**	501.1 (48.8)	99.5 (22.3)	50.0 (16.1)	10.0 (7.2)
Pedigree50	**5000 (800)**	503.21 (69.1)	99.8 (31.8)	49.9 (23.0)	10.0 (10.3)

Table S2. 3 Summary Statistics for HUNT and SardiNIA HDL Phenotype

		Male						Female					
	Sample Size	**N**	**Mean**	**Median**	**Min**	**Max**	**Age (mean, median)**	**N**	**Mean**	**Median**	**Min**	**Max**	**Age (mean, median)**
HUNT	5637	3717	47.9	46.4	19.3	116.0	62,61	1920	56.7	54.1	23.2	139.2	70,68
SardiNIA	5916	2506	58.7	57.1	21.3	147.7	42,44	3410	68.4	66.9	28.0	135.1	42,43

Table S2. 4 Count of Variants in HUNT and SardiNIA Exome Chip Data

	[0,0.01]	(0.01,0.05]	(0.05,1]	Total
HUNT	63,202	9,094	21,535	93,831
SardiNIA	41,986	9,699	25,143	76,828
Total				117,958[a]

a. This is the total number of sites that are polymorphic in the pooled HUNT and SardiNIA sample.

Table S2. 5 Count of Shared Variants in HUNT and SardiNIA Exome Chip Data

		HUNT			
		[0,0.01]	**(0.01,0.05]**	**(0.05,1]**	**Total**
SardiNIA	**[0,0.01]**	20,149	3,646	212	24,007
	(0.01,0.05]	2,817	3,631	1,534	7,982
	(0.05,1]	168	1,345	19,198	20,711
	Total	23,134	8,622	20,944	52,700[a]

a: This is the number of shared variants between HUNT and SardiNIA ExomeChip data.

Table S2. 6 Count of Non-Shared Variants in HUNT and SardiNIA Exome Chip Data

MAF	[0,0.01]	(0.01,0.05]	(0.05,0.5]	Total
HUNT	40,068	472	590	41,130[a]
SardiNIA	17,979	1,717	4,432	24,128[b]
Total				65,258[c]

a: this is the number of variants that are polymorphic in HUNT but not in SardiNIA.
b: this is the number of variants that are polymorphic in SardiNIA but not in HUNT.
c: the count of variants that are monomorphic in one study, but polymorphic in the other.

Table S2. 7 Time Usage and Key Features of Various Tools

	Run Time	Peak Memory Use	Kinship Options	Input Files	Implementation
famrvtest	1.5 hrs[a]	1.3G	Pedigree-based kinship or empirical kinship estimated from genotype within tool	VCF or Merlin PED/DAT file	C++ command line software
famskat	798 hrs[b]*	3.9G	Pedigree-based kinship generated by kinship R package	summarized genotype matrix	R function
pedgene	20 hrs*	4.4G	Pedigree-based kinship generated by kinship2 R package	summarized genotype matrix	R package
FFBSKAT	13 hrs[c]*	22G	Pedigree-based kinship generated by kinship2 R package or estimated from genotype by GenABEL	summarized genotype matrix	R package

All time collections were based on one CPU time analyzing 4000 genes of 1000 base-pair in a family sample of 10,000 individuals in 1000 families.

a. Time used for generating both burden and SKAT results starting from a VCF file.
b. famSKAT only allows a gene per run, thus the total run time is linear to number of genes analyzed. Time to analyze one gene is 2.6 minutes.
c. 9.9 hours were used to fit the linear mixed model under the null. 3.3 hours were used for calculating SKAT association statistics and p-values.

*. Time usage does not include summarizing raw data into usable genotype matrix.

CHAPTER 3: ASSOCIATION AND META-ANALYSIS METHODS FOR CHROMOSOME X

Introduction

The X chromosome contains 5% information of the human genome sequence, but contributes fewer GWAS findings than even chromosome 21 [Wise et al. 2013], which is considerably smaller. This is mostly because appropriate statistical methods and tools are sparse making analysis inconvenient, especially for family samples. Although low frequency and rare variant association analysis is becoming routine for GWAS, there are limited tools and methods for gene-level and rare variant association on the X chromosome.

Many work on X chromosome association methods for quantitative traits have been published. The work of Clayton et al. [Clayton 2008] proposed a score test for single variants on X chromosome in unrelated individuals.The XQTL approach [Zhang et al. 2009]uses a mixed model including separate random effects for X chromosome and the autosomes while modeling X-linked marker association as a fixed effect, which can be further decomposed into within family and between family contributions to overcome population stratification [Abecasis et al. 2000; Fulker et al. 1999]. However, this method is limited to nuclear families, and the decomposition naturally causes loss of power. MINX [Abecasis 2002], MERLIN in X, uses a variance component approach to model polygenic effects from both autosome and X chromosome and provides score and

likelihood ratio tests for X-linked marker associations, which allows analysis with small arbitrary pedigrees. We are also aware that much work has been done on qualitative traits X-linked marker associations in family samples [Chung et al. 2007; Clayton 2008; Thornton et al. 2012; Zhang et al. 2008; Zheng et al. 2007], and X-linked QTL linkage analysis methods have been extensively studied and implemented in tools that are now widely used [Abecasis et al. 2002; Almasy and Blangero 1998; Ekstrom 2004; Lange and Sobel 2006]. We extend this prior work, which focused on linkage analysis methods and also on the analysis of single variants. Here, our work focuses on gene-level association and meta-analysis methods for quantitative trait in families and also on enabling practical strategies for modeling and controlling for population structure. All our methods are naturally applicable to unrelated individuals – since these are simply a special sampling strategy where each family includes a single individual.

In this paper, we extend currently popular gene-level association and meta-analysis methods, SKAT, burden and VT tests, to X chromosome rare variants analysis. Our methods build upon the recent insight that gene-level association test statistics and meta-analysis statistics can be reconstructed from single-variant summary statistics [Feng et al. 2015; Liu et al. 2014]. Since most genes on X chromosome are down regulated through X-inactivation, such that only one "randomly" selected allele is expressed in cell or cell lineage [Chow et al. 2005], our methods are designed to account for this: basically, they assume that males hemizygous for an allele A will have the same phenotype as a female homozygous for A/A and that the phenotype of heterozygous females will be intermediate relative to the phenotype of opposite homozygotes. We evaluate type I error

and power of our approach using simulated data and 74 quantitative traits collected from SardiNIA study [Pilia et al. 2006].

Method

In this section, we describe a variance component model to handle relatedness among individuals and how to reconstruct gene-level association statistics from single variant summary statistics and covariance matrices, for both single study and meta analysis. Then, we summarize how we evaluated our method under the null and alternative hypothesis and the SardiNIA study dataset where we evaluated real life performance of our approach.

Variance Component Model

Given a sample of *n* individuals, we model quantitative trait values **y** as:

$$\mathbf{y} = \mathbf{X\beta} + \beta_{sex}\mathbf{Sex} + \mathbf{g} + \mathbf{g_X} + \boldsymbol{\epsilon}$$

Here, **Sex** is the indicator variable vector encoding sex for each sample, β_{sex} is the effect size for sex, **X** is the design matrix with relevant clinical covariates and intercept, **β** is the vector of effect sizes of the covariates, **g** and $\mathbf{g_X}$ are vectors of random effects modeling additive genetic effects from the autosomes and the X chromosome respectively, and **ϵ** is the vector of random error. We assume **ϵ** follows a multivariate normal distribution with mean **0** and covariance $\mathbf{I}\sigma_e^2$. We also assume the autosomal and X chromosome genetic effects **g** and $\mathbf{g_X}$ follow multivariate normal distributions with mean **0** and covariance $2\sigma_g^2\mathbf{K}$ and $2\sigma_{g_X}^2\mathbf{K_X}$, respectively. Matrices **K** and $\mathbf{K_X}$ summarizes kinship coefficients from the autosomes and X chromosome[Lange 1997]. σ_g^2 is a non-negative number

describing the autosomal polygenic variance contribution. σ^2_{gx} is a non-negative number that quantifies the X chromosome's polygenic contribution to trait variance. As systematically described in previous works [Kent et al. 2005; Zhang et al. 2009], assuming complete female X-inactivation, variance explained by X-linked polygenic effect in females (σ^2_{gx}) will be half of the variance explained in males ($2\sigma^2_{gx}$). In the overall sample, the variance explained by X-linked polygenes is $(2-r)\sigma^2_{gx}$ where r is the proportion of females in the sample, and total phenotypic variance equals $\sigma^2_g + (2-r)\sigma^2_{gx} + \sigma^2_e$. Mean and covariance of $\mathbf{y}$ are $E(\mathbf{y}) = \mathbf{X\beta}$ and $Cov(\mathbf{y}) = 2\sigma^2_g\mathbf{K} + 2\sigma^2_{gx}\mathbf{K_X} + \mathbf{I}\sigma^2_e$. Parameter estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma^2_g}$, $\widehat{\sigma^2_{gx}}$ and $\widehat{\sigma^2_e}$ are obtained using maximum likelihood. Although, in simple models of X-inactivation, males and females can have the same mean, we strongly recommend including sex as a covariate in the model to account for possible mean differences between the sexes.

Kinship coefficient is the probability of randomly drawing two chromosomes that are identical by decent (IBD) from a pair of individuals. We consider two approaches for estimating kinship matrices $\mathbf{K}$ and $\mathbf{K_X}$. One approach uses known pedigree structure to calculate the expected kinship, an alternative approach uses marker genotypes across the entire genome to calculate observed kinship. Since males inherit X chromosome only from their mothers, the values and algorithms for $\mathbf{K_X}$ are different from those for $\mathbf{K}$. For example, for non-inbred individuals, the self-kinship coefficients for are 1 (for males) and 0.5 (for females) on chromosome X; they are 0.5 (for both sexes) on autosomes. Estimating $\mathbf{K_X}$ between family members was fully discussed in [Kent et al. 2005; Lange 1997] and summarized in [Zhang et al. 2009].

Our alternative method estimates kinship using the Balding-Nicols empirical estimator [Astle and Balding 2009], $\widehat{\mathbf{K}} = \frac{1}{v}\sum_{i=1}^{v} \frac{(\mathbf{G_i}-2f_i\mathbf{1})(\mathbf{G_i}-2f_i\mathbf{1})^{\mathbf{T}}}{4f_i(1-f_i)}$ (here, v is the count of variants, $\mathbf{G_i}$ is a genotype vector where each element encodes the number of observed minor alleles in a particular individual, and f_i is the estimated minor allele frequency for the i^{th} variant). We estimate $\mathbf{K_X}$ using this equation, but coding female genotypes as 0,1, or 2 (depending on the number of minor alleles), and male genotypes as 0 or 2 (depending on presence or absence of the minor allele). This method allows us to quantify relatedness between apparently unrelated individuals and also allows for stochastic variation among pairs of individuals with the same degree of relatedness based on available pedigree data.

Association and Meta-analysis Methods

As previously described in our work [Feng et al. 2015], autosomal gene-level association test statistics can be reconstructed from single variant summary statistics and their covariance matrix. In this section, we apply the same idea to gene-level association and meta-analysis statistics for the X chromosome.

We first calculate use a score test for each variant [Chen and Abecasis 2007] and compute summary statistics and their covariance. The score statistic for the i^{th} variant is $U_i = (\mathbf{G_i} - \overline{\mathbf{G}}_\mathbf{i})^T\widehat{\mathbf{\Omega}}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$ where $\mathbf{G_i}$ is the genotype vector of this variant, $\widehat{\mathbf{\Omega}}$ is the estimated covariance matrix of phenotype $\mathbf{y}$, and $\widehat{\boldsymbol{\beta}}$ is the estimated fixed effect of clinical covariates. The covariance matrix of *m* score statistics of *m* variants in a gene is $\mathbf{V} =$

$(\mathbf{G}-\bar{\mathbf{G}})^{\mathrm{T}}(\hat{\boldsymbol{\Omega}}^{-1}-\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\Omega}}^{-1})(\mathbf{G}-\bar{\mathbf{G}})$ where $\mathbf{G}$ is the $n\times m$ genotype matrix. Under the null, score test of *i*th variant $\mathrm{T_i}=\frac{\mathrm{U_i}}{\sqrt{V_{ii}}}\sim\mathrm{N}(0,1)$ asymptotically.

Burden and Variable Threshold (VT) tests evaluate a similar model

$$\mathbf{y}=\beta_{\mathrm{sex}}\mathbf{Sex}+\mathbf{X\beta}+\gamma(\mathbf{G}-\bar{\mathbf{G}})\mathbf{w}+\mathbf{g}+\mathbf{g_X}+\boldsymbol{\varepsilon}$$

where γ is the fixed effect of the tested gene and null hypothesis is $\mathrm{H_0}:\gamma=0$. The burden score test statistic for an X-lined gene can be reconstructed from U and V as $\mathrm{T_{burden}}=\frac{\mathbf{w^T U}}{\sqrt{\mathbf{w^T V w}}}\sim\mathrm{N}(0,1)$ asymptotically. $\mathbf{w}$ is the vector of weights for the *m* variants in a gene. The VT test uses the maximum absolute burden score statistics over all possible frequency thresholds as the test statistic $\mathrm{T_{VT}}=\max_{\mathrm{F}}|\mathrm{T_{burden_F}}|$, and $\mathrm{T_{burden_F}}=\frac{\boldsymbol{\phi}_{\mathbf{F}}^{\mathbf{T}}\mathbf{U}}{\sqrt{\boldsymbol{\phi}_{\mathbf{F}}^{\mathbf{T}}\mathbf{V}\boldsymbol{\phi}_{\mathbf{F}}}}$ where $\boldsymbol{\phi}_{\mathbf{F}}^{\mathbf{T}}$ is a vector of 0s and 1s indicating if a variant is included by a specific frequency threshold Null distribution and p-value evaluation has been described in detail in [Lin and Tang 2011; Liu et al. 2014].

Sequence Kernel Association Test (SKAT) evaluates a different model

$$\mathbf{y}=\mathbf{X\beta}+\beta_{\mathrm{sex}}\mathbf{Sex}+\boldsymbol{\gamma}(\mathbf{G}-\bar{\mathbf{G}})+\mathbf{g}+\mathbf{g_X}+\boldsymbol{\varepsilon}$$

where γ_{i} is effect size of the *i*th variant in a gene and is randomly distributed with mean 0 and variance τw_i. The null hypothesis is $\mathrm{H_0}:\tau=0$. The SKAT statistic is $\mathrm{T_{SKAT}}=\mathbf{U^T W U}$ where $\mathbf{W}=\mathrm{diag}(\mathrm{w_1},\mathrm{w_2},\ldots,\mathrm{w_m})$ is a diagonal matrix of weights for each variant in a gene. Null distribution of $\mathrm{T_{SKAT}}$ and evaluation of significance was thoroughly described in [Chen et al. 2013; Wu et al. 2011].

To reconstruct gene-level meta-analysis statistics, we define single variant score statistics for meta-analysis as $U_{meta_i} = \sum_{k=1}^{s} U_{ik}$ and $V_{meta_{ij}} = \sum_{k=1}^{s} V_{ij,k}$, where *s* is the total number of studies, U_{ik} and $V_{ij,k}$ are score statistics and elements of covariance matrix from study *k* [Liu et al. 2014]. Then, gene-level association statistics can be established using the same method as described above for a single study.

Simulations

To evaluate our variance component model, we simulated quantitative traits under the null. We used three-generation pedigrees with a female to male ratio of 1:1 (see **Figure S1** for pedigree structure). Then the maximum likelihood estimates of variance components were averaged over 1,000 simulations and compared with the true generating values. To evaluate power, we simulated 10,000 haplotypes of 1,000 base-pair sequences, which is close to the length of an average protein coding sequence in humans. We simulated the haplotypes using ms [Hudson 2002] and a demographic model calibrated to mimic European population history [Adams and Hudson 2004; Novembre et al. 2008], and then we randomly selected 20% variants with frequency <0.01 as causal variants and defined effect sizes so that they altogether explained various amount of trait variance. We then randomly assigned founder haplotypes from the pool of haplotypes generated by ms, assigning one haplotype to each male founder and two to each female founder. We then successively sampled haplotypes for each descendant according to Mendel's Laws. When evaluating type I error, autosomal and X-linked polygenic effects, together with random error, were simulated. When evaluating power, causal gene effects were simulated on top of these.. We note autosomal heritability $h^2 = \frac{\sigma_g^2}{Var(y)}$, and X heritability $h_X^2 = \frac{\sigma_{gX}^2}{\sigma_g^2+\sigma_{gX}^2+\sigma_e^2}$.

SardiNIA Study

To evaluate our method in real data, we used exome chip data from the SardiNIA [Giorgio et al. 2014; Pilia et al. 2006] study, which genotyped 6,602 individuals. We analyzed 74 cardiovascular and personality quantitative traits, adjusted for age, sex and squared age, on 81,559 variants on exome chip where 1,543 variants were from chromosome X. Among X-linked variants, 817 had MAF<.01, 1,012 had MAF<.05. Among 79,980 autosomal variants, 44,557 had MAF<.01, and 53,742 had MAF<.05. Genotypes were called using the Illumina GenCall algorithm in combination with zCall V2.2. Detailed QC procedures can be found in [Giorgio et al. 2014].

Results

Accuracy of Heritability Estimates

To calibrate our variance component model, we simulated quantitative traits under the null where autosomal heritability were either 40% or none and X heritability varied from 5% to 40%, in samples of 1,000 individuals with a three-generation pedigree with 10 individuals in each family and sex ratio 1:1 (see **Figure S1**). We then averaged variance component estimates over 1,000 simulations for each setting. **Table 1** shows that when both autosomal and X chromosomal heritability are not zero, all 95% confidence intervals include true values. **Table 1** also shows that when true X heritability was zero, including X variance component in the model caused minimal false attribution of variance (1%) to chromosome X; this is expected since variance components are always estimated as positive. When the autosomal heritability was zero (**Table S1**), autosomal variance component captured a minimal amount of variance (<2% among all scenarios), and X heritability estimates were less than 1% smaller than true simulated values.

Type I error

To evaluate type I error, we simulated quantitative traits under the null in samples of 5,000 individuals in pedigree10 (see **Figure S1**) together with 2,000 genes under the null. **Figure 2** and **Figure S2** show that when X heritability was 10%, regardless autosome heritability was zero or 40%, QQ plots from all of our four gene-level tests were under control. **Figure S3** shows that if X heritability was 10% but we ignored the X variance component in the model, then QQ plot was really out of calibration under the null.

Power and Proportion of Females

Assuming complete X-inactivation in female, additive genetic effect from a causal gene from male samples has variance that is twice of that from female samples [Kent et al. 2005]. This leads to a hypothesis that including more males in a sample has larger association power to detect a single variant or a gene of the same allelic effect size in both male and female. To evaluate this hypothesis, we simulated quantitative traits of 5,000 unrelated individuals with different proportion of females in the sample and genes with 1,000 base-pair length and pre-defined causal variants (20% variants with MAF<.01) and effect sizes such that the gene explains 1% of trait variance in female samples, and performed burden test grouping variants with MAF<.01. **Figure 2** confirms that, although larger male proportion means less causal variants collected in a sample, more males in a sample leads to larger association power to detect a gene explaining the same amount of variance in females. In a more extreme situation, all-male sample has three times more power (26%) than all-female sample (5%) to detect a gene explaining 1% of trait variance in females. **Figure S4** shows similar power advantage in all-male sample when the variance explained by a gene in females varied from 0.5% to 2%.

Autosome vs. X Chromosome Power

In practice, if a causal gene has the same set of causal variants and effect sizes, the power should be identical if the gene is autosomal compared to if the gene is X-linked and the sample contains only female. This is because causal variants have the same chances to be sampled in both situations and following the same segregation pattern in families. However, if there are males in a sample, then power is expected to be larger if the gene is X-linked than if the gene is autosomal. As before, we simulated 10,000 haplotypes of a gene with 1,000 base-pairs and randomly selected 20% variants with MAF<.01 to be causal and calculated effect sizes such that the gene explains 0.5%-2% trait variance as if the gene was autosomal (the same amount of variance explained in females if the gene was X-linked). We simulated quantitative traits with 40% autosomal heritability and 10% X heritability in families with Pedigree10 (**Figure S1**) where half of the samples were male. **Figure 3** shows that there was much larger power to detect an X-linked gene than an autosomal gene explaining the same amount of variance. For example, when the gene explained 1% trait variance, power to detect an autosomal gene of such effect was 5% while power to detect such an X-linked gene was 12%. **Figure S5** shows similar power advantage over an X-linked gene with 0% autosomal heritability and 10% X heritability compared with an autosomal gene explaining the same amount of variance with 40% autosomal heritability.

SardiNIA results

To demonstrate our methods using real data, we analyzed exome chip data from SardiNIA study on multiple cardiovascular and personality quantitative traits using both single variant and gene-level tests. Empirical kinship matrices from autosome and X

chromosome were estimated from genotype of variants with MAF>0.05. Seventy traits have genomic controls between 0.95 and 1.05. There are four traits with genomic control (GC) greater than 1.05 where the largest is 1.08. Among all traits, X chromosome have substantial amount of contribution to trait variance in G6PD level and HbA1C, with X heritability 27% and 10% individually, whereas the others showed X heritability of <4%. 70 traits showed greater than zero autosomal heritability over a range of 5%-56%. **Figure 4** uses G6PD level as an exemplar trait showing that type I error and GC were under control. **Table 2** reports signal detected from single variant tests of 1,543 polymorphic sites from X chromosome from exome chip using significance level 6.13×10^{-7} according by Bonferroni correction with 81,566 variants tested. All signal belong to three variants where two of them were rare with frequency <.01, and these variants were associated with twelve traits. Except G6PD level, all other eleven traits have only one significant association. Conditional analysis conditioning on the top signal in G6PD level, which is a relatively common variant with frequency 0.08 and p-value 6.7×10^{-77}, showed strengthened significance from 1.3×10^{-24} to 4.4×10^{-26} for the other significant rare variant with frequency 0.9%. This shows that this rare variant with position 153774337 on X chromosome with major allele C and minor allele A is not a shadow of the common variant nearby with the strikingly small p-value. Condition upon this rare variant obtained a smaller p-value for the common variant at position 153762634 with major and minor alleles G and A also. The conditional analysis shows that these two variants are likely to be independent signals instead of being shadows of each other.

Gene-level tests of all quantitative traits detected the association of gene *G6PD* with G6PD level using significance level 2.8×10^{-6} as shown in **Table 3**. All three variants included in gene-level tests had MAF<.01 and all had negative effects. Although the single variant at position 153774337 with frequency .009 and effect size of -0.8 largely drives the signal from gene-level tests (p=***), the burden p-value 7.8×10^{-29} is smaller than any of the single variant p-value. The other two extremely rare variants with frequency 3×10^{-4} and 4×10^{-4} both had larger effect sizes than the variant that dominates the signal. P-value from VT test was closer to burden p-value but a little less significant due to tradeoff of multiple testing. SKAT p-value was less significant compare to burden and VT tests in this example, since all variants had the same direction of effect. **Table 3** also shows that condition upon the common variant with the most significant p-value from single variant test at position 153762634 (see **Table 2**) leads to an even smaller p-value for all four gene-level tests. This is evidence showing that the gene-level test signal of *G6PD* is not due to shadowing the single common significant variant of strikingly low p-value nearby. Analyses that condition on the most significant variant at position 153774337 included in the gene with frequency 0.009 had a much less significant p-value, suggesting that the gene-level test signal was mostly driven by this variant. However, all four conditional p-values from four gene-level tests was smaller than either of the rare variant included in the test, which shows synergy of these two low-frequency variants contributing toward the significance of these gene-level tests.

Conclusion and Discussion

In this paper, as an extension of our previous work [Feng et al. 2015], we described gene-level association tests, including burden, variable threshold and sequence kernel

association tests that can be reconstructed from summary statistics and their covariance matrices from single-variant scan. This approach allows fast computation for multiple gene-level tests and flexible grouping strategies without analyzing raw data repeatedly. This also extends to powerful meta-analysis approach where raw data sharing is impossible. We also demonstrated that for an X-linked gene with same effects, more males in a sample have larger power; and there is larger power to detect an X-linked gene than an autosomal gene with same effect sizes.

As pointed out by [Kent et al. 2005], assuming complete X-inactivation, male variance is twice of female and male and female have equal mean. Using this simple relationship, we use a variance component model with three variance components accounting autosomal, X chromosomal, and non-shared environmental contributions. We also described that empirical relationship from X chromosome can be estimated using genotypes from X-linked common variants. This expands our methods to handling possible cryptic relatedness, relatedness when pedigree structure is not known, distant relatedness, and possible population structure and provide more calibrated results under the null.

However, like any other method, our approach has assumptions that can be violated such that the model is not valid. For example, residual correlation may still exit after modeling relatedness from autosome and X chromosome because of other causes of phenotypic similarity such as shared-environment. If we fail to take account variance contribution of these extra causes in the model, and they happen to be quite large, then the model that we proposed might be off calibration. In this sense, checking QQ plots for any evidence of

inflated or deflated type I error after analysis is always recommended before making any conclusions.

The assumption of complete X-inactivation might not always be true either. However, as pointed out in [Carrel and Willard 2005; Payer and Lee 2008], incomplete X-inactivation is often the basis of various degrees of female X chromosome anomaly. As discussed in [Ober et al. 2008], if the trait of interest is not about X chromosome abnormalities in female, then the assumption of complete X-inactivation is valid, but we recommend excluding samples with this anomalies from analysis when feasible. It is also worth to point out that, in the situation of incomplete X-inactivation, female and male might have different mean. In this case, sex should be included as a covariate in the model.

There are other assumptions that could be violated in real data analysis. For example, allele frequencies might be different in male and female due to many causes, such as heterogeneous ethnicity and sequencing or genotyping error. Many other reasons might also cause difference in mean between sexes, for example men and women differ naturally in average height, BMI and some personality traits. Thus, we recommend to always add sex as a covariate to account for the possible difference in mean. Equal effect sizes in male and female might also be violated due to regulatory difference between sexes in some traits. If there is previous knowledge showing that this is true, then the interaction between sex and genotype should be included in the model and sex should be included as a covariate.

Finally, although this paper mainly focuses on methods for family samples, population samples can be considered a special case of family sample and our methods and tools are feasible to samples of unrelated individuals. Our method has been implemented in RAREMETAL, a freely available software that supports multiple platforms including Linux, MAC, and Windows. The documentation, source code and executable can be downloaded in the following:

http://genome.sph.umich.edu/wiki/RAREMETAL_Documentation

Tables

Table 3. 1 Autosomal and X Chromosomal Heritability Estimates Under the Null

Polygenic Source	True Heritability	Mean Estimates	SE	Lower 95% Confidence Level	Upper 95% Confidence Level
Chr. X	0.05	0.05	0.0007	0.050	0.052
Autosome	0.40	0.40	0.0012	0.396	0.401
Chr. X	0.15	0.15	0.0008	0.149	0.152
Autosome	0.40	0.40	0.0012	0.396	0.401
Chr X	0.10	0.10	0.0008	0.099	0.103
Autosome	0.40	0.40	0.0013	0.397	0.402
Chr. X	0.25	0.25	0.0008	0.249	0.252
Autosome	0.40	0.40	0.0012	0.396	0.401
Chr. X	0.20	0.20	0.0008	0.199	0.203
Autosome	0.40	0.40	0.0012	0.396	0.401
Chr. X	0.35	0.35	0.0009	0.349	0.352
Autosome	0.40	0.40	0.0012	0.396	0.401
Chr. X	0.30	0.30	0.0008	0.299	0.302
Autosome	0.40	0.40	0.0012	0.396	0.401
Chr. X	0.40	0.40	0.0009	0.399	0.402
Autosome	0.40	0.40	0.0012	0.397	0.402
Chr. X	0.00	0.01	0.0004	0.007	0.009
Autosome	0.40	0.39	0.0009	0.388	0.392

Each grid represents a simulation setting with a certain combination of autosomal heritability (h^2) and X chromosomal heritability (h_x^2). Results were summarized from 1,000 simulations.

Table 3. 2 Single Variant Hits on Chromosome X from SardiNIA Quantitative Traits Association Tests

Trait	GC*	Chr:Pos:A1:A2[$]	N	MAF[@]	Effect Size (SD)[#]	p-value	Conditional p-value**
G6PD Level	0.99	X:153762634:G:A	6613	0.08	-1.19	$6.7x10^{-77}$	$2.3x10^{-78}$
G6PD Level	0.99	X:153774337:C:A	6613	0.009	-0.79	$1.3x10^{-24}$	$4.4x10^{-26}$
HbA1C	1.03	X:153762634:G:A	6434	0.08	-0.74	$1.1x10^{-58}$	
RBC	1.03	X:153762634:G:A	6724	0.08	-0.29	$9.9x10^{-20}$	
Bilirubin, total	0.96	X:153762634:G:A	6198	0.08	0.32	$2.4x10^{-18}$	
MCV	1.02	X:153762634:G:A	6724	0.08	0.27	$9.9x10^{-15}$	
Serum Iron	0.96	X:153762634:G:A	6769	0.08	0.22	$1.6x10^{-12}$	
MCH	1.03	X:153762634:G:A	6724	0.08	0.23	$3.4x10^{-11}$	
BMI	1.05	X:153036439:A:G	6770	0.0003	2.87	$1.0x10^{-8}$	
Waist	1.06	X:153036439:A:G	6770	0.0003	2.53	$6.6x10^{-8}$	
Weight	1.03	X:153036439:A:G	6770	0.0003	2.68	$4.4x10^{-8}$	
Bilirubin, fractionated	1.01	X:153762634:G:A	6198	0.08	0.18	$2.2x10^{-8}$	
Ferritin	0.98	X:153762634:G:A	4936	0.09	0.17	$2.3x10^{-7}$	

$6.13x10^{-7}$ was used as p-value cutoff for hits. Table is sorted by smallest p-value and trait.
*GC represents genomic control.
[$]A1 is the major allele, and A2 is the minor allele.
[@]MAF is the minor allele frequency and was calculated from founders. [#]Effect sizes of minor allele were measured in standard deviations.
**If there are more than one significant variants for a trait, then conditional analysis was performed. Only G6PD level has two hits. The first p-value was from conditional analysis conditioning on X:153774337:C:A, and the second p-value was p-value from conditioning on X:153762634:G:A.

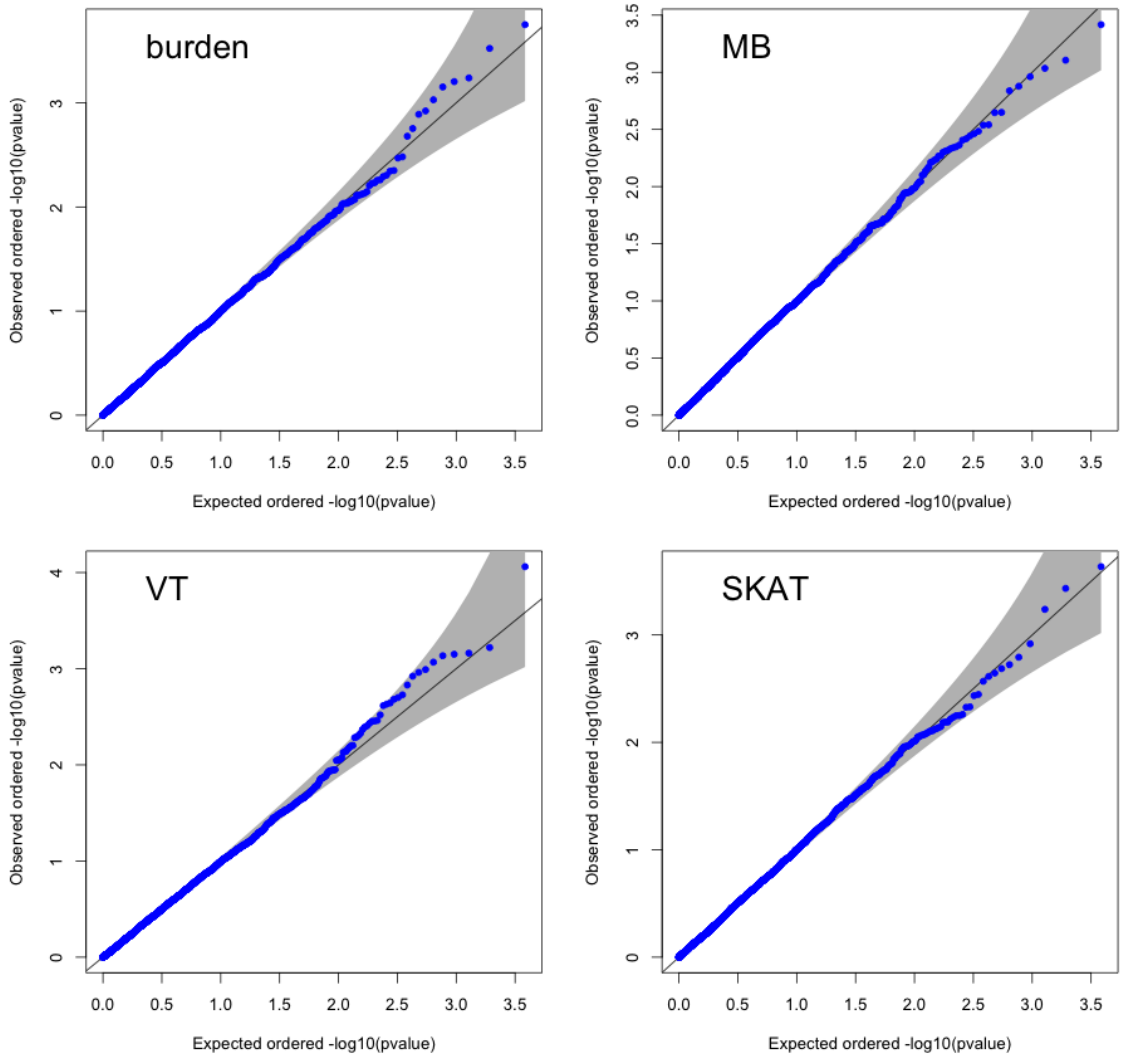
Table 3. 3 Gene-level Association and Conditional Analysis of SardiNIA G6PD Level

Gene	Variants Included[$]	MAF	Effect Size (SD)*	P-value*	Gene-level Test p-value			
					Burden	MB	SKAT	VT
Unconditioned								
G6PD	X:153761811:C:G	0.0004	-1.24	$1.0x10^{-4}$	$7.8x10^{-29}$	$1.3x10^{-21}$	$4.0x10^{-25}$	$2.3x10^{-28}$
	X:153764217:C:T	0.0003	-1.18	$9.8x10^{-3}$				
	X:153774337:C:A	0.009	-0.79	$1.3x10^{-24}$				
Condition on X:153762634:G:A[&]								
G6PD	X:153761811:C:G	0.0004	-1.22	$1.2x10^{-4}$	$2.3x10^{-30}$	$2.1x10^{-22}$	$1.3x10^{-26}$	$6.8x10^{-30}$
	X:153764217:C:T	0.0003	-1.22	$8.0x10^{-3}$				
	X:153774337:C:A	0.009	-0.81	$4.4x10^{-26}$				
Condition on X:153774337:C:A								
G6PD	X:153761811:C:G	0.0004	-1.21	$1.4x10^{-4}$	$3.2x10^{-6}$	$3.2x10^{-6}$	$2.2x10^{-5}$	$5.6x10^{-6}$
	X:153764217:C:T	0.0003	-1.24	$6.7x10^{-3}$				
	X:153774337:C:A	0.009						

*These are results from single variant tests. Effect sizes are for minor alleles.
[$]Variants included are in the format of Chr:Pos:MajorAllele:MinorAllele.
[&]X:153762634:G:A is the variant that is most significant from single variant test for G6PD (See **Table 3.2**) with frequency 0.08.
G6PD levels were quantile normalized before association analysis.

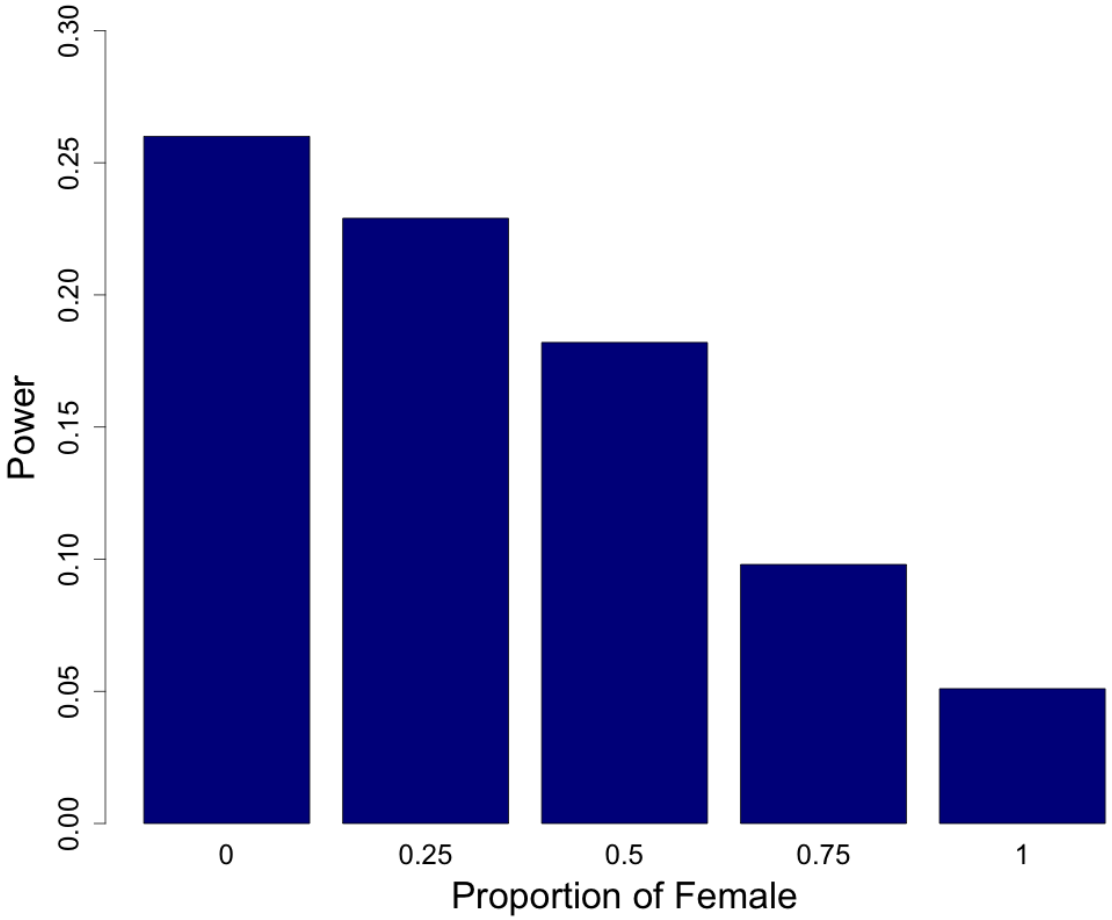
Figures

Figure 3. 1 QQ Plots under Null with both Autosomal and X Contribution to Variance



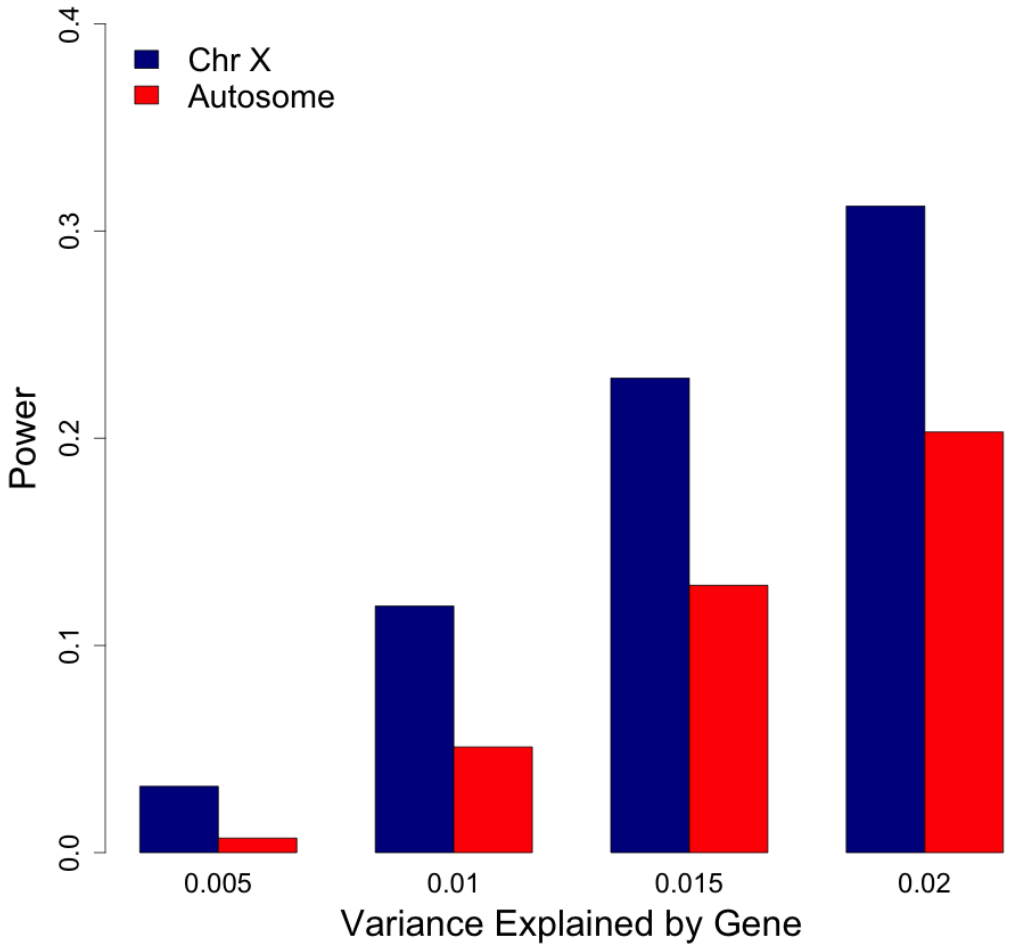
Quantitative traits were simulated in 5,000 individuals with pedigree10 (see **Figure S3.1** for structure) under the null model. Autosomal heritability was 40% and X heritability was 10%. 2,000 genes were simulated starting from founder haplotypes and then gene-drop to children along pedigrees. Variant with frequency <0.01 were grouped for four gene-level tests and then QQ plots were generated based on the obtained p-values.

Figure 3. 2 Power and Proportion of Females in a Sample



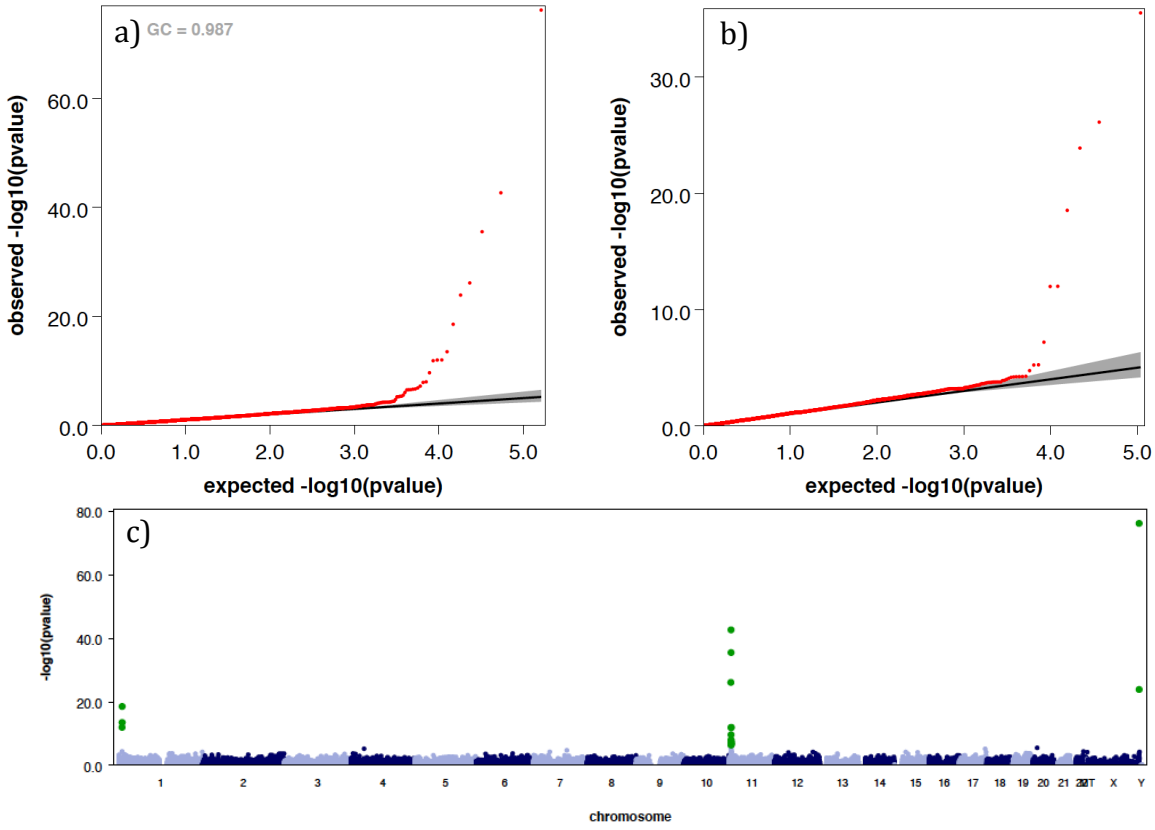
Power was averaged over 1,000 simulations. 10,000 haplotypes for a gene with 1k base-pair were simulated. Then 20% variant with frequency <0.01 were assigned to be causal. Effect sizes of causal variants were calculated such that causal variants altogether explain 1% trait variance in females. Then quantitative traits of 5,000 unrelated individuals were simulated based on causal effects and with various proportion of female samples. Simple burden test was performed on each simulated data set.

Figure 3. 3 Power to Detect an Autosomal Gene and an X-linked Gene



10,000 haplotypes with1k base-pairs in length were simulated. 20% variants with frequency <0.01 were selected to be causal and effect sizes calculated such that the gene explains 0.5%-2% proportion of trait variance in females. Founder haplotypes of 5,000 individuals with pedigree structure shown in **Figure S3.1** were randomly selected from the pool of haplotypes. Children haplotypes were gene-dropped from parents haplotypes. Quantitative traits were simulated with 40% autosomal heritability (for autosomal associations) and 10% X heritability plus 40% autosomal heritability (for X-linked associations) together with causal gene effects.

Figure 3. 4 QQ and Manhattan Plots of SardiNIA G6PD Trait Association from Exome Chip Single Variant Test



Autosomal and X-chromosomal empirical kinship matrices were estimated from common variants with frequency >0.05 on Exome Chip, and used for fitting the variance component model for this analysis. Autosomal and X-chromosomal heritability was estimated to be 28% and 27% individually. Genomic control was 0.99.

a) is the QQ plot for all variants.
b) is the QQ plot for variants with frequency <5%.
c) is the Manhattan plot of all variants. Green dots are variants that passed p-value threshold using Bonferroni correction, which was 6.13×10^{-7}.

Supplementary

Figure S3. 1 Pedigree Structure Used in Simulation

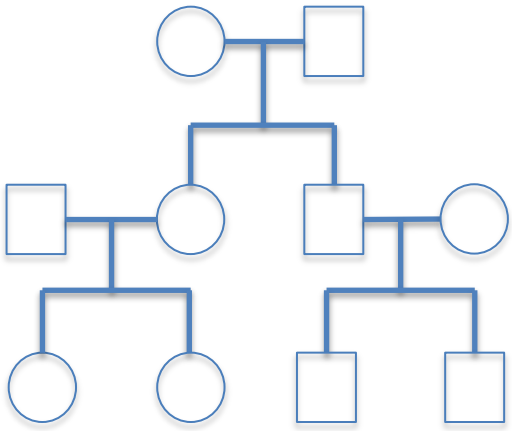
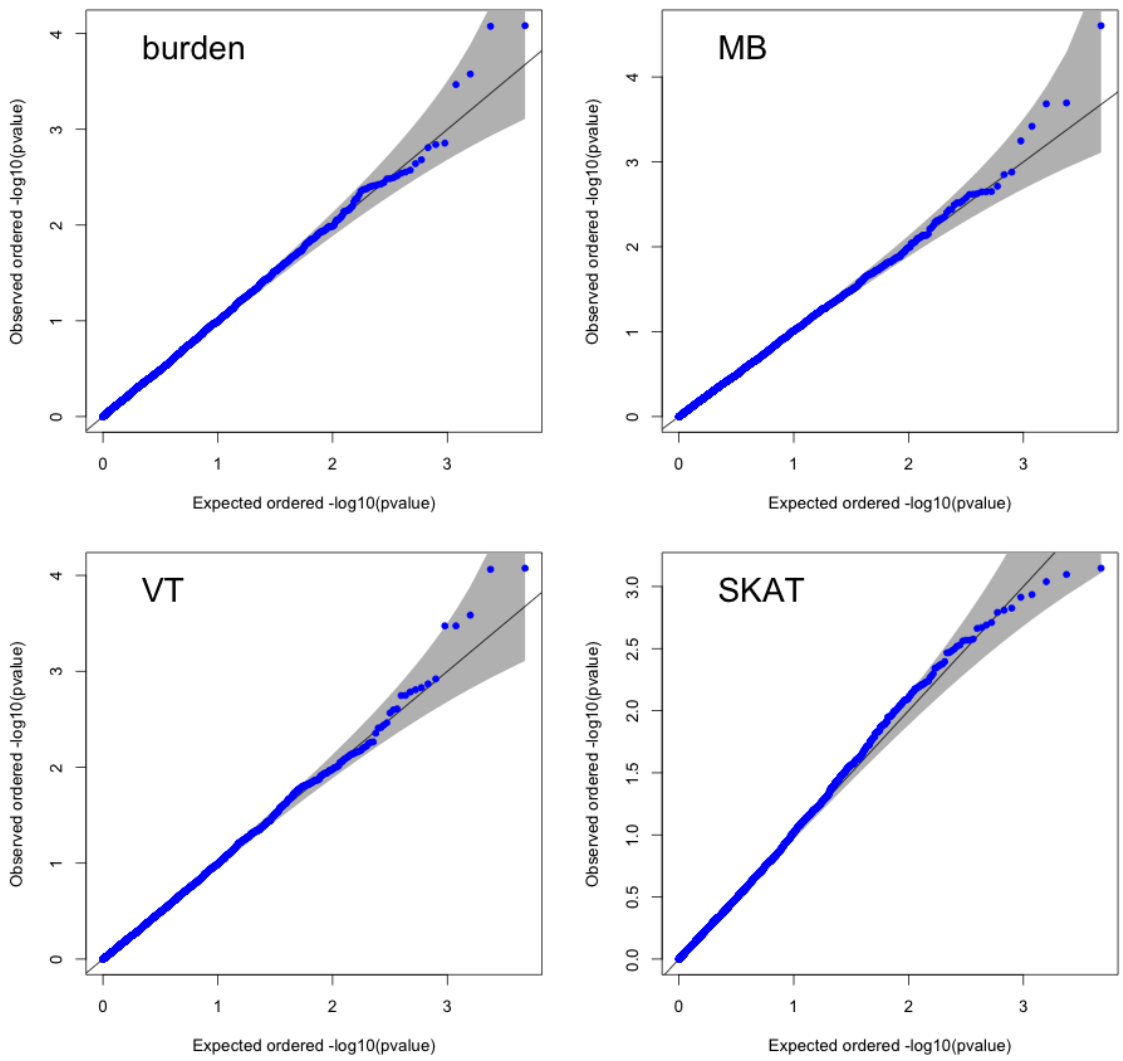
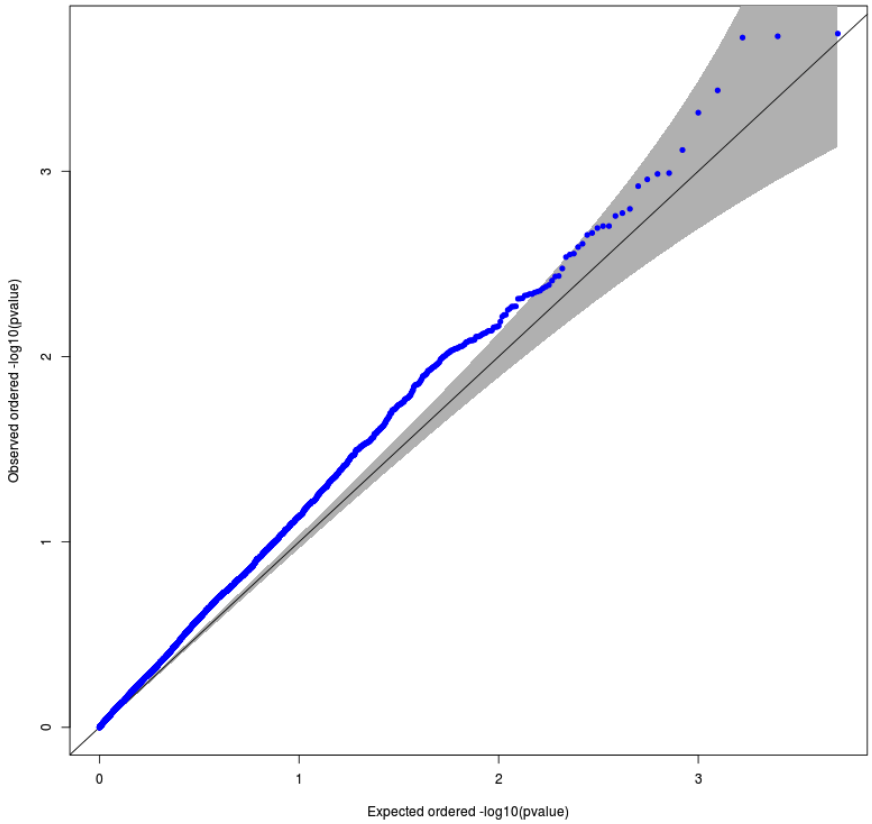


Figure S3. 2 QQ Plots under Null with no Autosomal Contribution to Variance



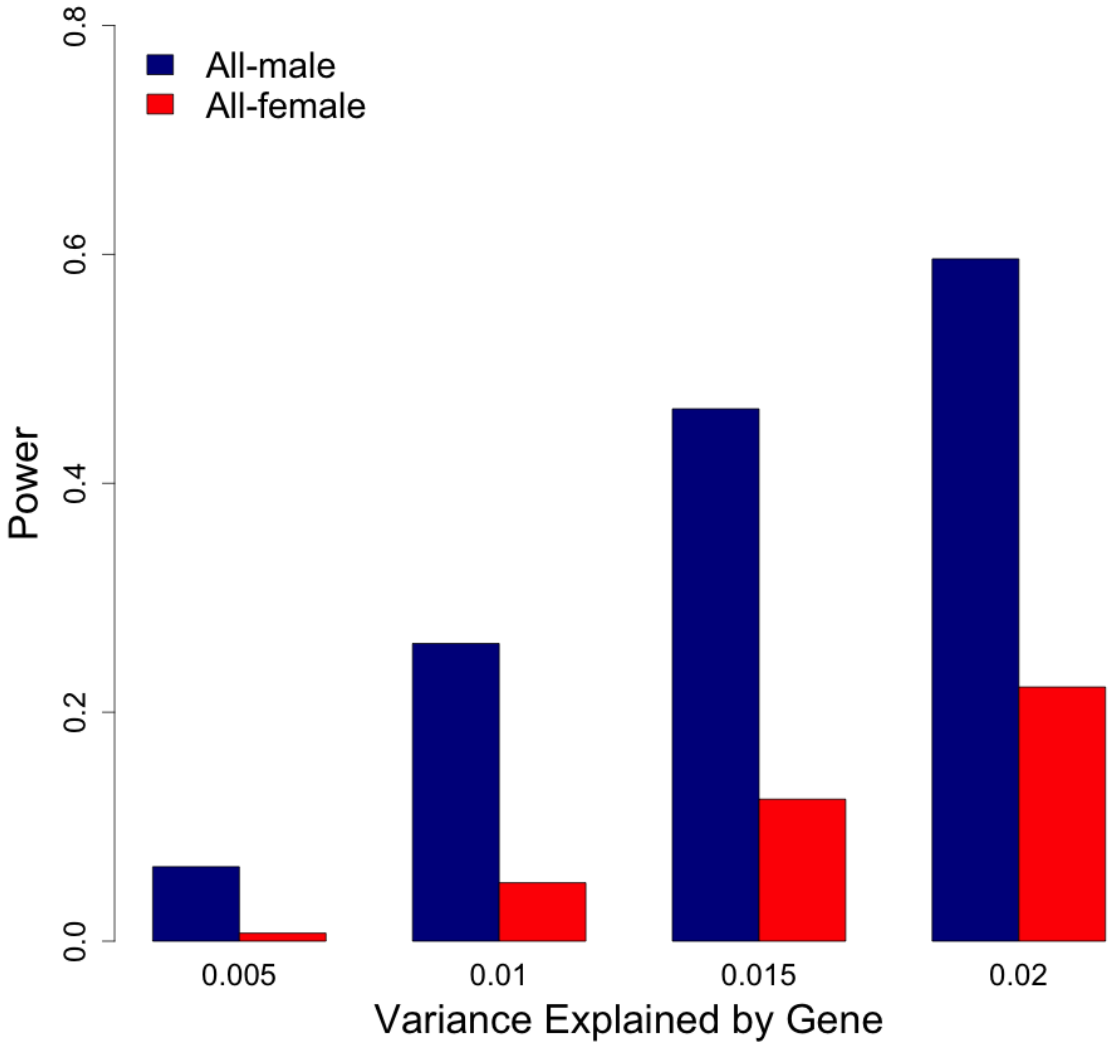
Quantitative traits were simulated in 5,000 individuals with pedigree10 (see **Figure S3.1** for structure) under the null model. Autosomal heritability was 0% and X heritability was 10%. 2,000 genes were simulated starting from founder haplotypes and then gene-drop to children along pedigrees. Variant with frequency <0.01 were grouped for four gene-level tests and then QQ plots were generated based on the obtained p-values.

Figure S3. 3 QQ plot when Ignoring X Variance Component



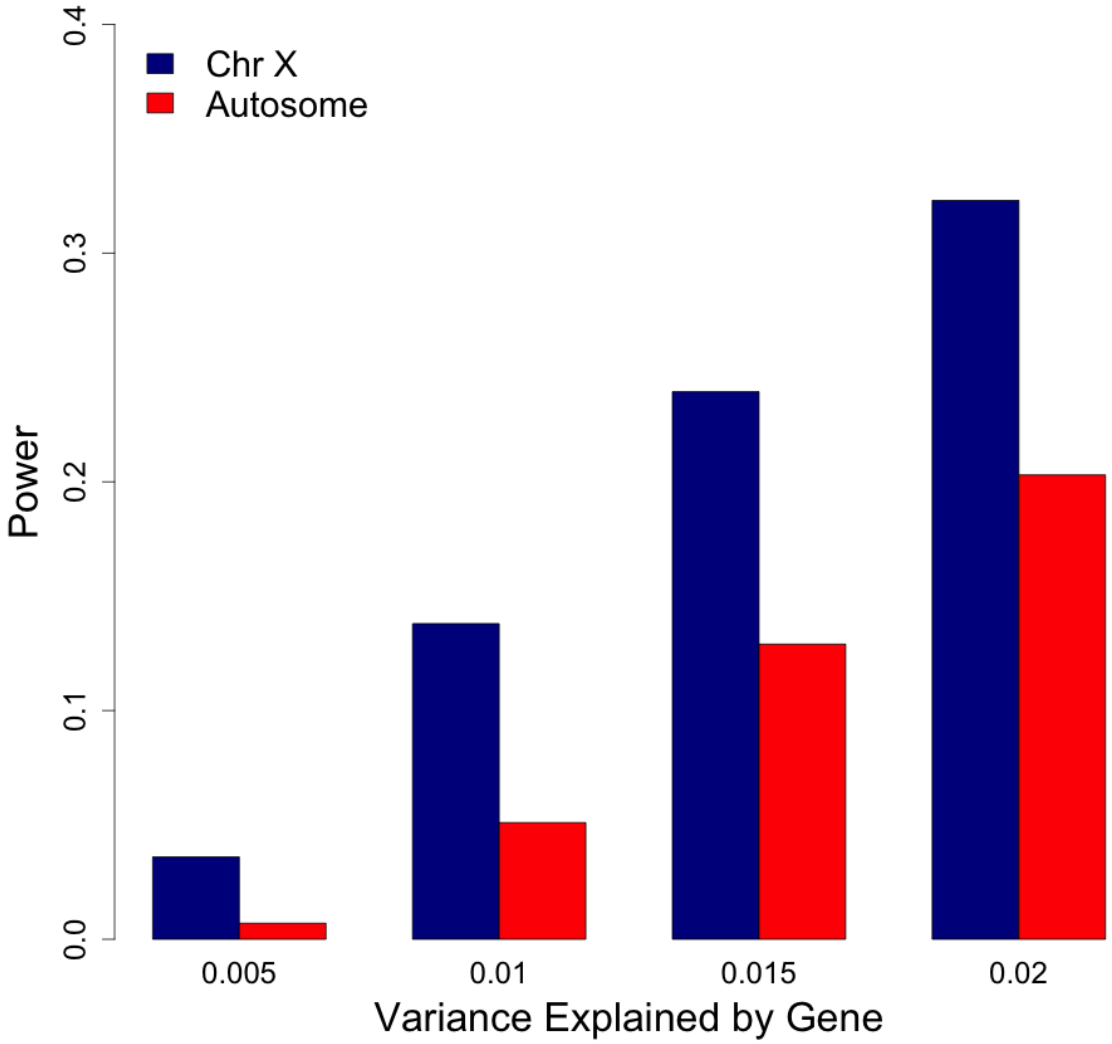
The QQ plot was generated from associations of 2,000 genes under the null where autosomal and X heritability was 40% and 10%, individually. A simple variance component model was fit where X variance component was ignored.

Figure S3. 4 Power Comparison in All-male vs. All-female Samples to Detect a Gene Explaining Various Proportion of Trait Variance in Females



Power was averaged over 1,000 simulations. 10,000 haplotypes for a gene with 1k base-pair were simulated. Then 20% variant with frequency <0.01 were assigned to be causal. Effect sizes of causal variants were calculated such that causal variants altogether explain various trait variance in females, ranging from 0.5% to 2%. Then quantitative traits of 5,000 unrelated individuals were simulated based on causal effects and were either all male or all female. Simple burden test was performed on each simulated data set.

Figure S3. 5 Power to Detect an Autosomal Gene and an X-linked Gene



10,000 haplotypes with1k base-pairs in length were simulated. 20% variants with frequency <0.01 were selected to be causal and effect sizes calculated such that the gene explains 0.5%-2% proportion of trait variance in females. Founder haplotypes of 5,000 individuals with pedigree structure shown in **Figure S3.1** were randomly selected from the pool of haplotypes. Children haplotypes were gene-dropped from parents haplotypes. Quantitative traits were simulated with 40% autosomal heritability (for autosomal associations) and 10% X heritability plus 0% autosomal heritability (for X-linked associations) together with causal gene effects.

Table S3. 1 Autosomal and X Chromosomal Heritability Estimates Under the Null with no Autosomal Polygenic Effects

Polygenic Source	True Heritability	Mean Estimates	SE	Lower 95% Confidence Level	Upper 95% Confidence Level
Chr X	0.05	0.04	0.0007	0.041	0.044
Autosome	0.00	0.01	0.0008	0.013	0.016
Chr X	0.15	0.14	0.0007	0.141	0.143
Autosome	0.00	0.01	0.0007	0.013	0.016
Chr X	0.10	0.09	0.0007	0.091	0.093
Autosome	0.00	0.01	0.0007	0.013	0.016
Chr X	0.25	0.24	0.0007	0.239	0.242
Autosome	0.00	0.02	0.0008	0.014	0.017
Chr X	0.20	0.19	0.0007	0.190	0.193
Autosome	0.00	0.02	0.0008	0.014	0.017
Chr X	0.35	0.34	0.0008	0.339	0.342
Autosome	0.00	0.02	0.0008	0.014	0.017
Chr X	0.30	0.29	0.0008	0.289	0.292
Autosome	0.00	0.02	0.0008	0.014	0.017
Chr X	0.40	0.39	0.0008	0.388	0.391
Autosome	0.00	0.02	0.0008	0.015	0.018
Chr X	0.00	0.00	0.0002	0.003	0.004
Autosome	0.00	0.01	0.0004	0.005	0.006

Each grid represents a simulation setting with a certain combination of autosomal heritability (h^2) and X chromosomal heritability (h_x^2). Results were summarized from 1,000 simulations.

CHAPTER 4: IDENTIFYING TRAIT-ASSOCIATED RARE VARIANTS BEFORE SEQUENCING

Introduction

Most non-synonymous, splice altering, and protein truncating variants are very rare[Abecasis et al. 2010; Abecasis et al. 2012; Marth et al. 2011; Nelson et al. 2012].Because these variants have clear functional consequence, discovery of association between a medically relevant trait and these variants can provide clear insights about disease biology.[Boucas et al. 2013; Raychaudhuri et al. 2011; Zhan et al. 2013] Next-generation sequencing has been accelerating the discovery of these rare trait-associated variants, but remains expensive.

Prioritizing individuals and families to sequence from existing samples remains important, and often relies on the simple identification of phenotypic extremes.[Ahituv et al. 2007; Cohen et al. 2004; Cohen et al. 2006; Sanna et al. 2011] This approach can enrich samples for rare alleles with large effects, and facilitates discovery of trait-associated rare variants.[Guey et al. 2011; Kryukov et al. 2009; Sanna et al. 2011; Van Gestel et al. 2000] Another popular strategy is to study family samples, because pedigrees make it simpler to identify multiple copies of trait associated rare variants. Many publications successfully discussed selection strategies typically for sibship samples to gain association power.[Abecasis et al. 2001b; Kwan et al. 2009; Risch and Zhang 1995; Risch and Zhang 1996] In addition, in families, it may often be possible to

sequence a subset of individuals and, through genotype imputation, propagate this information to their relatives.[Cheung et al. 2014] This strategy is implemented in tools such as PRIMUS,[Staples et al. 2013] which selects a set of maximally unrelated samples; GIGI-Pick,[Cheung et al. 2014] which optimizes genotype imputation in candidate regions based on pedigree structure and genotype; and ExomePicks,[Abecasis 2011] which selects individuals from large families to maximize ability to estimate and impute rare variant haplotypes.

Here, we propose a new approach for prioritizing individuals for sequencing. RAREFY selects individuals and families that are likely to carry trait-associated rare variants. The approach models background polygenic effects using a variance component model and can adjust for covariates and the effects of known variants. Our approach relies on the intuition that by examining the segregation of phenotypes across the entire pedigree, it should be possible to better prioritize individuals for sequencing than by examining individual phenotypes alone. For example, intuitively, we might expect that an individual with an extreme phenotype who has a parent and a child who are also phenotypically extreme in the same direction may be more likely to carry a variant of large effect than an individual whose parent and child are phenotypically average (**Figure S4.1**). Our approach considers all possible variant segregation patterns and prioritizes individuals and families that seem more likely to carry a trait associated rare variant.

We show that for a fixed sequencing effort our approach has more power to discover and associate more trait-associated rare variants than methods that focus on individuals phenotypic extremes alone.

Method

In this section, we first explain the variance component model that is used to handle familial relationship. Then, we describe our approach for small pedigrees, where we enumerate and evaluate all possible genotype configurations, conditional on family structure. This is followed by a description of a Markov Chain Monte Carlo (MCMC) approach for larger pedigrees to evaluate the most likely genotype configurations, conditional on family structure and available phenotypes. The reason that the MCMC method is introduced is that for large pedigrees computational cost grows exponentially using the former approach, but the MCMC approach avoids the enumerations thus makes our method computationally feasible for large pedigrees. Finally, we describe simulations and the data set used as a real data example.

Modeling Familial Relatedness

Our first step is to calculate residuals of quantitative traits, taking account familial relationship and key covariates. We assume the usual linear model[Falconer and Mackay 1996]

$$\mathrm{E}(\mathbf{y_i}) = \boldsymbol{\mu} + \mathbf{X_i}\boldsymbol{\beta_x},$$

(Equation 1)

where $\mathbf{y_i}$ is the phenotype vector for n individuals in the i^{th} family, $\boldsymbol{\mu}$ is the vector of population mean, $\mathbf{X_i}$ is the design matrix, and $\boldsymbol{\beta_x}$ is the vector of covariate effects. Then $\mathbf{y_i}$ follows multivariate normal distribution with mean $\boldsymbol{\mu} + \mathbf{X_i}\boldsymbol{\beta_x}$ and covariance matrix $\boldsymbol{\Omega_i} = \sigma_g^2\mathbf{K_i} + \sigma_e^2\mathbf{I}$, where $\mathbf{K_i}$ is the kinship matrix,[Lange 1997] σ_g^2 is the genetic variance component, and σ_e^2 is residual environmental variance. Parameters $\boldsymbol{\mu}$, $\boldsymbol{\beta_x}$, σ_g^2, and σ_e^2 are estimated using maximum likelihood. For convenience, we define $\widehat{\boldsymbol{\Omega_i}} = \widehat{\sigma_g^2}\mathbf{K_i} + \widehat{\sigma_e^2}\mathbf{I}$, as the estimated covariance matrix of $\mathbf{y_i}$ and the trait residuals vector as $\widetilde{\mathbf{y}_\mathbf{i}} = \mathbf{y_i} - \widehat{\boldsymbol{\mu}} - \mathbf{X_i}\widehat{\boldsymbol{\beta}_\mathbf{x}}$.

Method for Smaller Pedigrees

Our approach calculates the expected number of copies of a rare allele in each individual or family for a putative rare variant with large effect size, based on observed phenotype and estimated variance components and fixed effects. Let $\mathbf{g_i} = (g_{i1}, g_{i2}, \ldots, g_{in})$, be a vector of genotypes coded 0, 1, or 2 copies of the rare allele. Define a random variable S_{ij} as the count of trait-associated rare alleles in the j^{th} individual of the i^{th} family. Then the expectation of S_{ij} given the residual vector $\widetilde{\mathbf{y}_\mathbf{i}}$ is

$$E\left(S_{ij}\middle|\widetilde{\mathbf{y}_\mathbf{i}}\right) = \sum_{\mathbf{g_i}} g_{ij} P(\mathbf{g_i}|\widetilde{\mathbf{y}_\mathbf{i}}) = \frac{\sum_{\mathbf{g_i}} g_{ij} P(\widetilde{\mathbf{y}_\mathbf{i}}|\mathbf{g_i}) P(\mathbf{g_i})}{\sum_{\mathbf{g_i}} P(\widetilde{\mathbf{y}_\mathbf{i}}|\mathbf{g_i}) P(\mathbf{g_i})}$$

(**Equation 2**)

where the summations range over all possible genotype configurations. $P(\mathbf{g_i})$ is the probability of observing the genotype configuration $\mathbf{g_i}$ in a pedigree. $P(\widetilde{\mathbf{y}_\mathbf{i}}|\mathbf{g_i})$ is the conditional probability of residuals given a specific genotype configuration $\mathbf{g_i}$.

To calculate $P(\widetilde{\mathbf{y_i}}|\mathbf{g_i})$ in equation (2), we let β_g be the postulated additive effect size of a trait-associated variant with frequency *p*. Then $\widetilde{\mathbf{y_i}}|\mathbf{g_i}$ follows the multivariate normal distribution with mean $\beta_g\mathbf{g_i}$ and covariance $\widetilde{\mathbf{\Omega_i}} = \widetilde{\sigma_g^2}\mathbf{K_i} + \widehat{\sigma_e^2}\mathbf{I}$, in which $\widetilde{\sigma_g^2} = \widehat{\sigma_g^2} - 2p(1-p)\beta_g^2$. We write the conditional likelihood as

$$P(\widetilde{\mathbf{y_i}}|\mathbf{g_i}) = (2\pi)^{-\frac{n}{2}}\left|\widetilde{\mathbf{\Omega_i}}\right|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\left(\widetilde{\mathbf{y_i}} - \beta_g\mathbf{g_i}\right)^T\widetilde{\mathbf{\Omega_i}}^{-1}\left(\widetilde{\mathbf{y_i}} - \beta_g\mathbf{g_i}\right)\right\}$$

(**Equation 3**)

It is straightforward that the expected count of the rare alleles for the i^{th} family can be calculated using

$$E(S_i|\widetilde{\mathbf{y_i}}) = \sum_j E\left(S_{ij}\middle|\widetilde{\mathbf{y_i}}\right).$$

(**Equation 4**)

For convenience, we name $E(S_i|\widetilde{\mathbf{y_i}})$ as the RAREFY-Family score and $E\left(S_{ij}\middle|\widetilde{\mathbf{y_i}}\right)$ as the RAREFY-Individual score.

As described, our method requires enumerating all possible genotype configurations in a pedigree and calculating the likelihood for each. The complexity of this calculation is exponential with family size *n*. In our implementation, exhaustive enumeration is only feasible for pedigrees with <25 individuals. In the next section, we describe a MCMC

approach for larger pedigrees. In our example dataset, we apply the MCMC approach to pedigrees including as many as 1,453 members.

MCMC for Larger Pedigrees

For large pedigrees, we use a Metropolis-Hastings[Hastings 1970; Metropolis et al. 1953] algorithm to sample from the posterior distribution of $P(\mathbf{g_i}|\tilde{\mathbf{y}}_\mathbf{i})$. For a family with *n* individuals, we consider each of 3^n possible genotype vectors as a potential state for a Markov chain. The Markov chain starts by randomly assigning founder genotypes and then randomly propagating founder alleles to offspring according to Mendel's laws. After the initial likelihood is calculated, new states are proposed by randomly selecting an individual to update in the genotype vector. The probability of a state at the t^{th} iteration is

$$L^{(t)} = P(\mathbf{g}_\mathbf{i}^{(\mathbf{t})}|\tilde{\mathbf{y}}_\mathbf{i}) \propto P(\tilde{\mathbf{y}}_\mathbf{i}|\mathbf{g}_\mathbf{i}^{(\mathbf{t})})P(\mathbf{g}_\mathbf{i}^{(\mathbf{t})}),$$

where $P(\tilde{\mathbf{y}}_\mathbf{i}|\mathbf{g}_\mathbf{i}^{(\mathbf{t})})$ is calculated from **Equation 3**. If the updated genotype vector is inconsistent with Mendelian inheritance, then $P(\tilde{\mathbf{y}}_\mathbf{i}|\mathbf{g}_\mathbf{i}^{(\mathbf{t})})$ is zero.

After convergence (typically millions of iterations), allowing burn-in period and thinning, we estimate the posterior mean of the genotype vector $\mathbf{g_i}|\tilde{\mathbf{y}}_\mathbf{i}$ of the i^{th} family. Then RAREFY-Individual score $E(S_{ij}|\tilde{\mathbf{y}}_\mathbf{i})$ can be obtained from the posterior mean of $\mathbf{g_i}|\tilde{\mathbf{y}}_\mathbf{i}$. RAREFY-Family score $E(S_i|\tilde{\mathbf{y}}_\mathbf{i})$ can then be calculated using **Equation 4**.

Simulations

We considered three types of pedigree structures (See **Figure S4.2**) in simulations: a nuclear family with three siblings and extended 3-generation pedigrees with 10 and 15 individuals, respectively. To simulate genotypes, we first simulated haplotypes for founders, using a population genetic model implemented in ms[Hudson 2002] to simulate 1,000 base-pair sequences, and then used gene-dropping to propagate these to other individuals in the pedigree. Causal single variants were simulated with effect size 1 or 2 trait standard deviations, and causal genes were simulated to explain 0.1-1.0% trait variance. In causal genes 20% variants with frequency <1% were selected to be causal and each assigned the same amount of trait variance (resulting in a different effect size for each variant). Total heritability for simulated quantitative traits was 40%, including polygenic background effects.

Empirical Significance Level and Power

We evaluated association power empirically using a family-based score test.[Chen and Abecasis 2007; Feng et al. 2014] 1,000 simulations were used to obtain power of each setting. To mimic the process of real studies, we first used phenotypes from the entire sample to estimate variance component and fixed effect parameters and to select individuals for "sequencing". In the association analysis stage, only genotypes of these selected individuals contributed to association tests. Because normality is usually violated and sample sizes were often small after selection, we estimated association p-values empirically using 300,000 permutations. In each permutation, we first simulated founder haplotypes for the variant using estimated founder allele frequency in selected samples,

and then used gene-dropping to simulate offspring haplotypes. The Besag-Clifford stopping rule[J. and P. 1991] was used to approximate the p-value with faster computation. Power was calculated as the proportion of the original 1,000 datasets that showed evidence of association at empirical significance level of 10^{-5}, due to limitation of computational cost from massive amount of permutation tests. This could limit our insight of the performance of our method at lower significance levels..

Selection Methods for Comparison

To evaluate the strengths and weaknesses of our approach, we compared RAREFY-Individual and RAREFY-Family methods with two other popular approaches: Extreme Phenotype and Extreme Unrelated Phenotype strategies. All methods used residuals generated after adjusting for study-specific covariates and known variants in linear models. The Extreme Phenotype method selects individuals with extremely high or low residuals. The Extreme Unrelated Phenotype method picks unrelated individuals with extremely low or high residuals, based on a modified implementation of the PRIMUS approach for selecting a maximum set of unrelated individuals with extreme phenotypes. We compared both discovery power (the number of the trait associated allele among selected individuals) and association power.

Exemplar Dataset

To evaluate the performance of our approach we also reanalyzed LDL-cholesterol levels in sample of individuals from the Lanusei valley in Sardinia (see **Table S4.2** for descriptive statistics), adjusting for age, sex, and squared age as covariates. Quantitative traits were quantile normalized before analysis but covariates were not transformed. The

SardiNIA sample includes 6,602 genotyped individuals (9,720 individuals including non-genotyped founders and ancestors). The largest pedigree includes 1,453 individuals spanning five generations. Genotypes were called using the Illumina GenCall algorithm in combination with zCall V2.2. Detailed QC procedures can be found in Pistis et al.[Giorgio et al. 2014]

Results

In this section, we compare the power to detect trait-associated rare variants and genes using RAREFY and alternative approaches. We also explore the impact of misspecified parameters on RAREFY analyses. Finally, we apply RAREFY to a study of LDL cholesterol in the isolated population of SardiNIA.

Discovery Power

To evaluate power for a single trait-associated rare variant, we simulated family samples of various sizes (N=2,500~20,000) and configurations (see **Figure S4.2**) and a causal variant with minor allele frequency 0.001 with effect size 1 or 2 trait standard deviations. We then compared the number of causal alleles captured using the RAREFY, Extreme Phenotype, Extreme Unrelated Phenotype, and random selection strategies.

Figure 4.1 shows that the RAREFY-Individual and RAREFY-Family methods always provide largest discover power for trait-associated rare alleles. Random selection provided the least power for variant discovery. Selecting extreme individuals without regard to their relatedness (the Extreme Phenotype approach) was better than selecting unrelated individuals with extreme phenotypes (the Extreme Unrelated Phenotype

approach) but still performed less well than RAREFY. For example, for a variant with frequency 0.001 and effect size 1 trait standard deviation, sequencing 200 individuals from the original sample of 20,0000 individuals using RAREFY captured approximately the same number of causal alleles (~4 copies) as sequencing 2,000 individuals selected randomly. Similarly, sequencing 500 individuals by RAREFY was as efficient as sequencing 1,000 or 1,800 individuals selected by Extreme Phenotype or Extreme Unrelated Phenotype strategies. When effect sizes are larger (2 trait standard deviations), RAREFY performed even better (see **Figure 4.1B**). For example, sequencing 500 individuals selected by RAREFY had the same discovery power for causal alleles (~25 captured) as sequencing 1,800 individuals by Extreme Phenotype method, and as sequencing > 2,000 individuals by other strategies.

Figure 4.1 also shows substantial enrichment of rare causal alleles in selected samples using RAREFY. For example, in **Figure 4.1B**, applying RAREFY-Individual to a sample of 20,000 individuals (and on average of 40 trait associated variants), prioritizing 1,000 individuals captured 28.2 causal alleles (versus 20.8 selecting phenotypic extremes alone) with effect size 2 trait standard deviations. In other words, sequencing 5% individuals captured 70.5% rare alleles, increasing the frequency of trait-associated variants from 0.1% in the original sample to 1.4% in the selected sample, a 14-fold increase.

Figure 4.2 shows that, with a fixed number of individuals to sequence, selecting from larger samples improves the value of selected individuals. For example, for a causal variant with frequency 0.001 and an original sample with 2,500 individuals, sequencing

1,000 captured 4.6 causal alleles (almost all of the 5 copies in the original sample). When the sample of phenotyped individuals increased to 20,000, the same sequencing effort (1,000 individuals sequenced) captured 28.2 causal alleles. Thus enriched allele frequency in two selected samples of 1,000 individuals was 0.2% and 1.4%, respectively.

Complex diseases are likely to be affected by multiple variants.[Willer et al. 2008] To explore this situation, we simulated a 1,000 base pair sequence where 20% of variants with frequency <0.01 were trait-increasing and, altogether, explained 1% of trait variance. In this setting, each variant has a different effect size and the model used by RAREFY to analyze the data is misspecified. Again, both RAREFY-Family and RAREFY-Individual scores were able to prioritize more causal alleles for "sequencing" (**Figure S4.3**), and the RAREFY-Individual method was able to pick the most causal alleles. Since variants with opposite effects can reside in the same gene,[Abifadel et al. 2003] we simulated genes where half of causal variants were trait-increasing and the remainder were trait-decreasing. **Figure S4.4** shows that, among the methods examined, RAREFY provided the best power compare to others in this situation.

RAREFY methods performed well regardless of pedigree size. **Figure S4.3-S5 s**how that pedigree structure and size does not affect discovery power of RAREFY-Individual method. In **Figure S4.3-S5**, Discovery power for RAREFY-Family method decreased with increasing family size, but remained second best.

Impact of Model Misspecification

Since RAREFY analyses requires postulating a frequency and effect size for trait-associated variants, we evaluated the impact of misspecifying these parameters. **Figure S4.6** shows that choice of MAF has almost no effect on discovery power of RAREFY when a true MAF ranged from $5x10^{-4}$ and 0.01 is misspecified between $5x10^{-4}$ and 0.05, and effect size is specified correctly. **Figure S4.7** shows that RAREFY continues to outperform selection strategies based on individual phenotypic extremes when true effect size ranged from 0.5 and 1.5 was misspecified as between 0.25 and 2.5, and MAF was specified correctly. However, parameter settings closer to the true (and typically unknown) effect size produced better discovery power. With the expectation that most interesting trait-associated variants for complex traits will have effect sizes greater than 0.5 standard deviations, together with simulation results shown in **Figure S4.6,** we choose MAF 0.001 and effect size 1 as default parameter values in RAREFY. All simulation results shown in previous sections were based on this default parameter setting.

Association Power

To evaluate power of association analysis in samples prioritized by RAREFY, we performed single variant association score test[Chen and Abecasis 2007] on selected samples evaluating p-value empirically using 300,000 gene-drops per simulated sample. **Figure 4.3** and **Figure 4.4** shows that, among the methods we considered, RAREFY-Family and RAREFY-Individual provides the largest and second-largest power for detecting a single trait-associated variant and random selection provides the least power. For example, using significance level 10^{-5}, **Figure 4.3** shows that sequencing 400

individuals or more from a sample of 20,000 individuals provides 89.9% and 82.8% power by RAREFY-Family and RAREFY-Individual; however, to obtain power >80%, Extreme Phenotype based selection required sequencing 800 individuals or more. Sequencing 2,000 individuals selected randomly only provided 22% power. **Figure 4.4** shows that, with fixed sequencing effort, selecting from a larger sample provides larger association power. For example, by RAREFY-Family, selecting 1,000 from 5,000, 10,000, and 15,000 individuals to sequence provides 48.1%, 78.9%, and 91.7% power respectively. Obtaining 80% power by selecting 1,000 individuals with Extreme Phenotypes required >18,000 phenotyped individuals.

Analysis of Exemplar Data Set

To evaluate RAREFY in real data, we analyzed LDL cholesterol levels in 6,602 individuals from the Lanusei valley in Sardinia. This is a relatively isolated population and includes many families, small and large[Pilia et al. 2006]. Age, sex and square of age were used as clinical covariates (see **Table S4.1** for descriptive statistics). Known associated variants (See **Table S4.2** for list of variants) were also used as covariates to obtain better discover power by RAREFY. We then evaluated our ability to identify carriers of rare variant V578A[Sanna et al. 2011] in *LDLR*, which is unique in SardiNIA and has frequency 0.005 (61 copies in the sample) and effect size 23.7mg/dl (0.63 standard deviation). To show the impact of misspecification of parameters to RAREFY power, we used various parameter combinations to run RAREFY. **Figure 4.5** shows that RAREFY typically provided higher power than methods based on selecting phenotypic extremes, even after parameter misspecification. For example, sequencing 1,000

individual by RAREFY using the default parameter settings (MAF=0.001, effect=1), RAREFY captured 26 copies of the rare allele, but sequencing Extreme Phenotype individuals captured 19. Sequencing 2,000 individuals (<1/3 of the entire sample), captured more than half of the rare alleles (32 copies). **Figure 4.5** also shows that RAREFY prioritizes carriers even when only small number of individuals are sequenced. For example, the top 100 individuals prioritized by RAREFY include 11 copies of V578, but those selected based on phenotype extremes alone include only 3. **Figure S4.9** and **S10** show that including known variants as covariates boost RAREFY power but provide only a limited benefit when selecting phenotypic extremes.

Tool and Computational Performance

We implemented our method in a C++. RAREFY is a command line tool that uses Merlin[Abecasis et al. 2002] format input files and can prioritize carriers of trait-increasing and trait-decreasing rare variants in small or large pedigrees. RAREFY rapidly handles most small pedigrees and supports parallel computing for samples including larger pedigrees. A RAREFY analysis of a sample of 20,000 individuals in 4,000 nuclear families of size five, searching for both trait-increasing and trait-decreasing variants, required 10.95 seconds on a single CPU. A RAREFY analysis of a family with ~1,500 individuals using MCMC with 50,000,000 iterations in five chains takes 8 hours using 5 CPUs.

Discussion

We describe a new approach to prioritize individuals and families that carry trait-associated rare alleles. Using simulation and real data analysis, our approach greatly

outperforms selection based on extreme phenotypes alone. RAREFY is able to handle both families and unrelated samples.

Our RAREFY-Individual method is able to capture more associated rare alleles than RAREFY-Family method, but our RAREFY-Family approach provided greater power for association analyses in our simulations. Part of the explanation, is that the RAREFY-Family method typically results in samples that include more diverse sets of phenotypes – including family members who are unlikely to carry trait associated rare variants but which help estimate phenotypic values for non-carriers.

Adjusting for previously associated variants is helpful in guiding searches for new trait-associated rare variants. Working in combination, these variants can account for extreme phenotypes in many individuals. We suggest genotyping samples at known loci and using these genotypes to adjust phenotype residuals. When direct genotyping is not feasible, existing array data together with genotype imputation procedure can also be useful.[Cheung et al. 2014] We do caution that adjusting for previously associated common variants could reduce the chance identifying rare causal variants that are in linkage disequilibrium with these.

RAREFY could be improved further by specifying a distribution for effect size, instead of a fixed value. Normal distribution has been widely used for distribution of effect sizes, but an appropriate prior distribution of parameters to specify the variance of the distribution of effect sizes should be carefully evaluated. For random effect size,

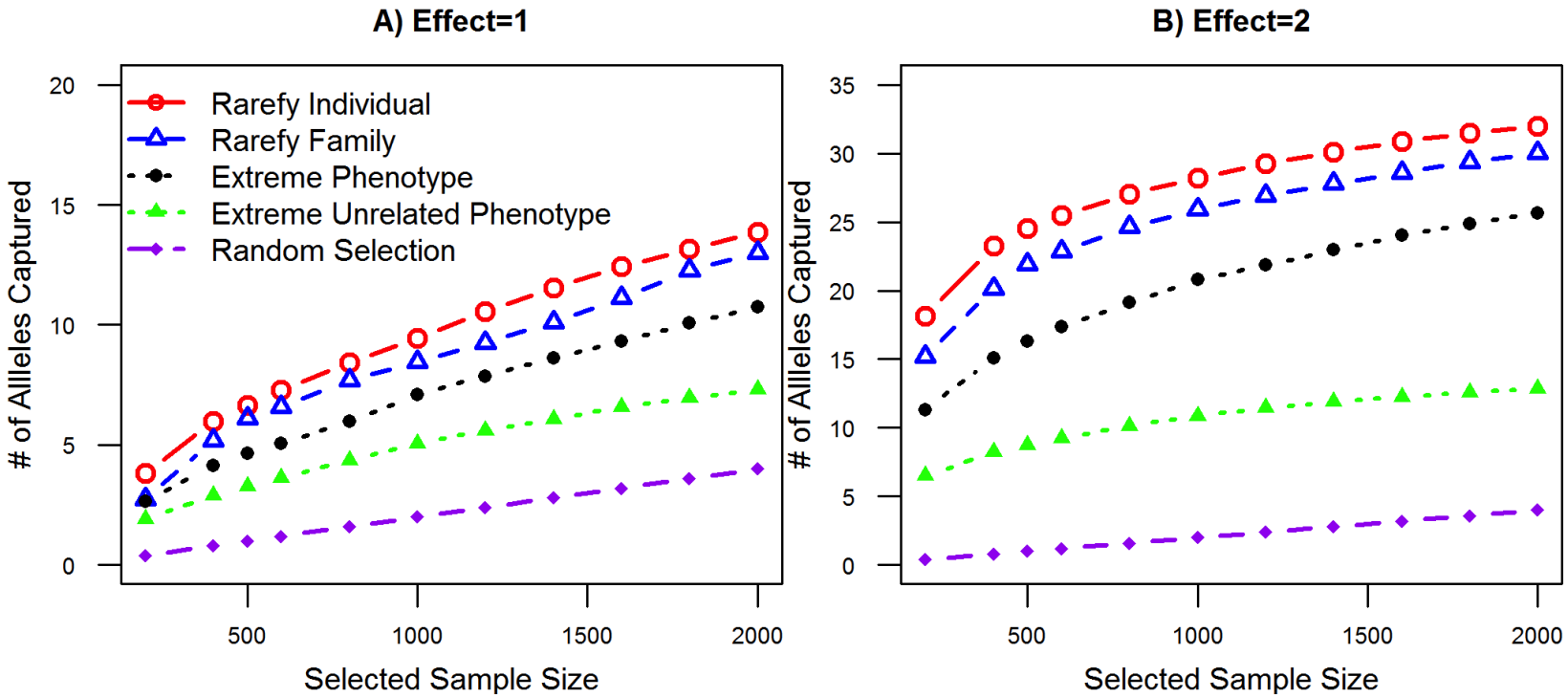
computational cost could also increase exponentially, depends on the distribution specified. Another improvement could be optimizing association test statistics, instead of discovery power.

We suggest selecting individuals and families for both trait-increasing and trait-decreasing rare variants, because both gain-of-function and loss-of-function rare variants are typically of interest[Abifadel et al. 2003]. After individuals and families are prioritized and sequenced and candidate rare variant are identified, genotyping the entire sample for these variants may help confirm and extend potential discoveries.

In summary, RAREFY provides powerful solutions, fast computation, and command-line tool to prioritize families and unrelated individuals to sequence among phenotyped individuals.

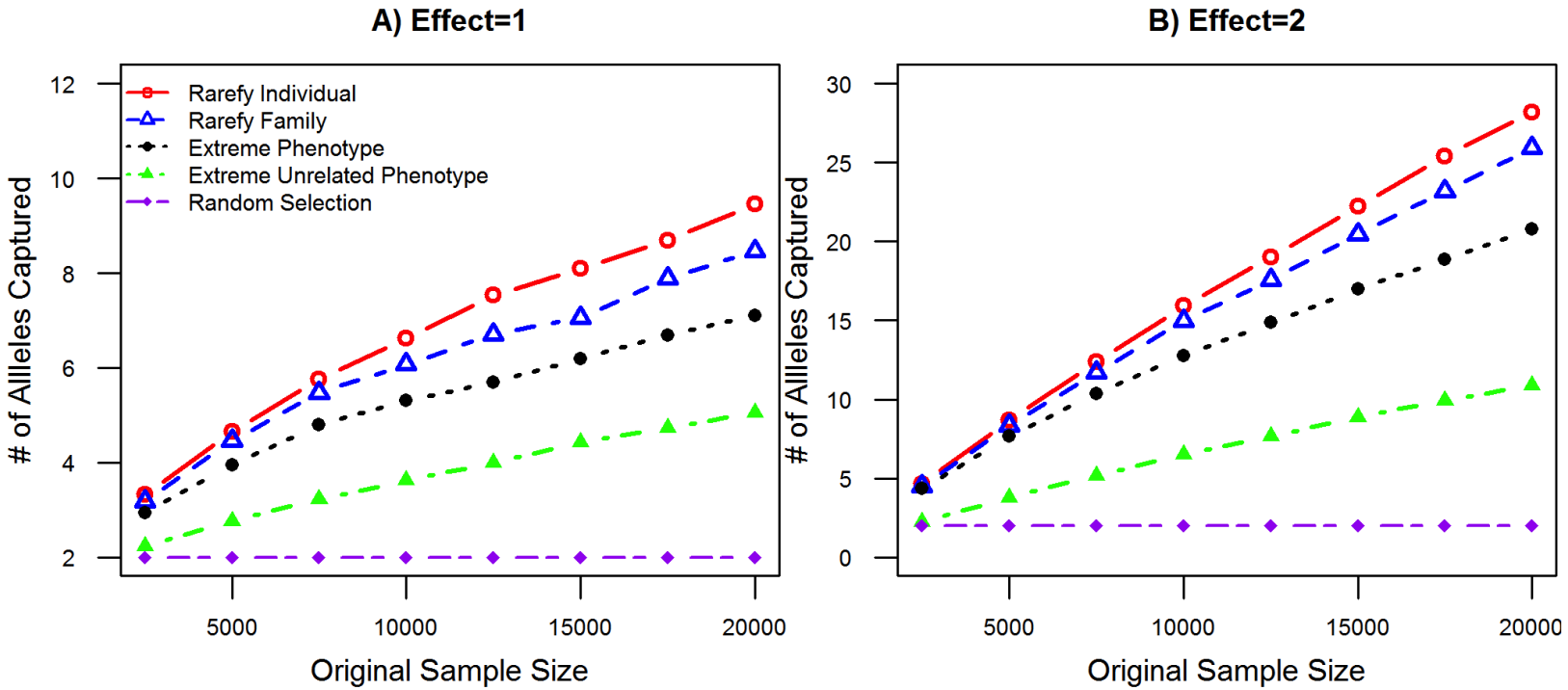
Figures

Figure 4. 1 Number of Alleles Captured for a Rare Variant (MAF=0.001) Sequencing 100-2,000 Individuals from a Sample of 20,000 Individuals



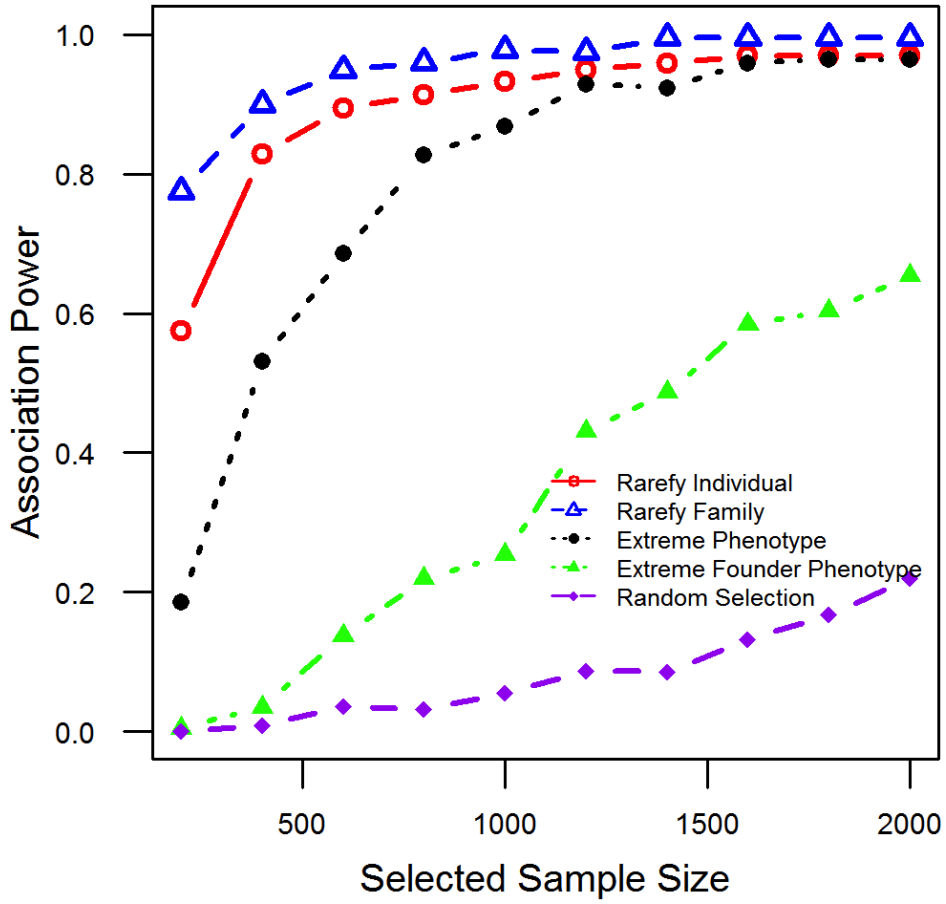
A single variant with frequency 0.001 with effect size = 1 or 2 trait standard deviation was simulated in founders and gene-dropped to children in original samples of 20,000 individuals, with Pedigree5 (see **Figure S4.1**) family structure.

Figure 4. 2 Number of Alleles Captured for a Rare Variant (MAF=0.001) Sequencing 1,000 Individuals from Original Samples of Various Sizes



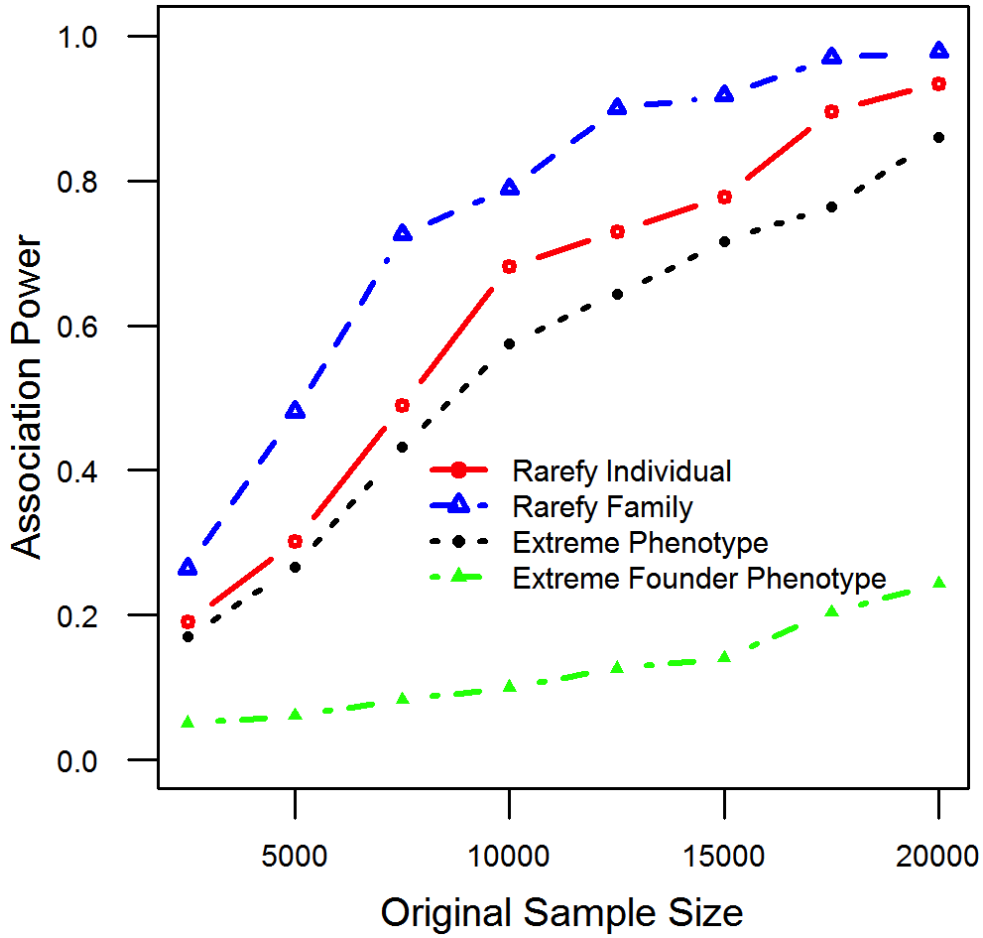
A single variant with frequency 0.001 with effect size = 1 or 2 trait standard deviation was simulated in founders and gene-dropped to children in original samples of various sizes (2,500~20,000), with Pedigree5 (see **Figure S4.1**) family structure.

Figure 4. 3 Association Power of Sequencing 200-2,000 Individuals from 20,000 Individuals



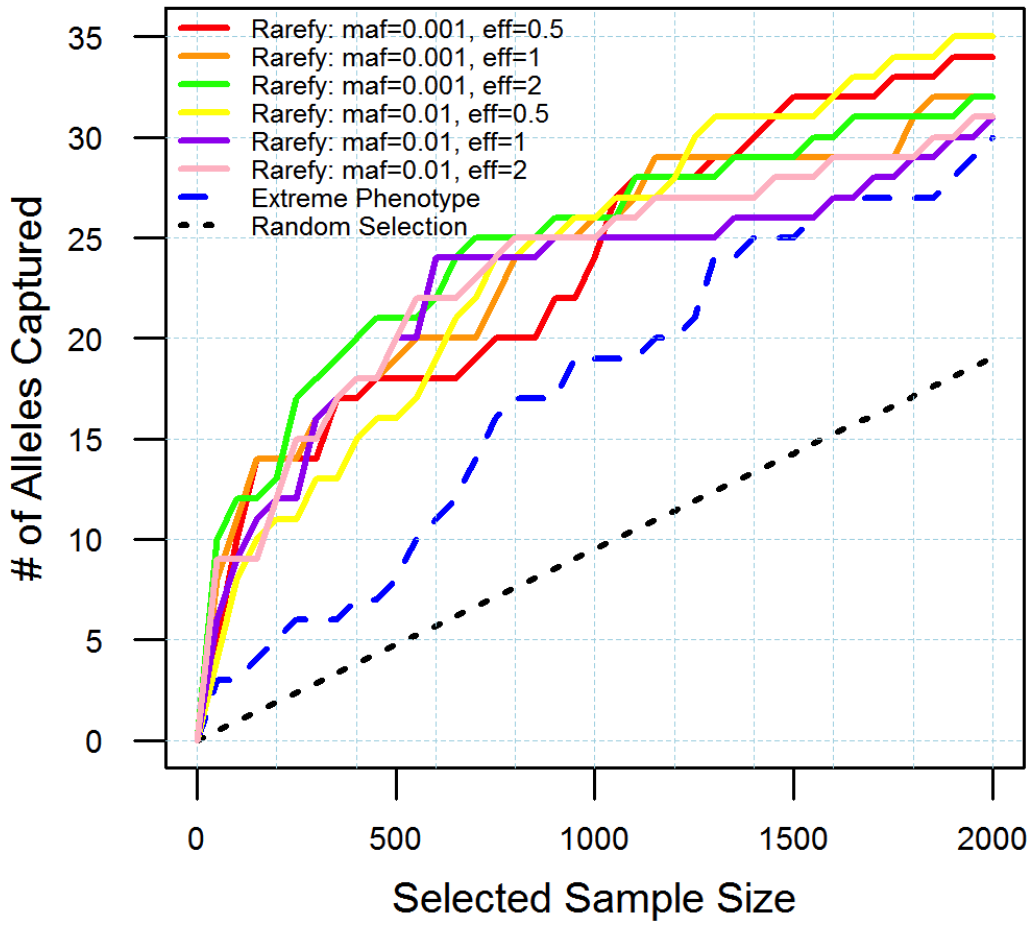
200-2,000 Individuals were selected from 20,000 individuals, using various selection strategies. 300,000 permutations and significance level 10^{-5} were used to obtain association power. All phenotyped samples were used to fit variance component model. Un-selected samples were set to have genotypes missing.

Figure 4. 4 Association Power of Sequencing 1,000 Individuals from Samples of Various Sizes by Different Strategies



1,000 Individuals were selected from original samples with 2,500-20,000 individuals, using various selection strategies. 300,000 permutations and significance level 10^{-5} were used to obtain association power. All phenotyped samples were used to fit variance component model. Un-selected samples were set to have genotypes missing.

Figure 4. 5 Power to Discover Rare Allele of V578* from SardiNIA Sample Using RAREFY-Individual and Extreme Phenotype Strategies



50-2,000 individuals were selected for sequencing from SardiNIA sample of 6410 phenotyped individuals and covariates adjusted. Adjusted LDL values (40 individuals taking cholesterol-lowering drugs were added 40 to their LDL values) were used as phenotype. LDL values were inverse-normalized before fitting linear mixed model. Covariates were age, sex, and squared age, together with known variants (See **Table S4.2** for list of known variants adjusted).
*V578A is a rare variant on chr19, position 11227562 in LDLR with minor allele frequency 0.005 and effect size 0.63 standard deviation, which is unique in SardiNIA sample. It has total allele count 61 in the sample.

Supplementary

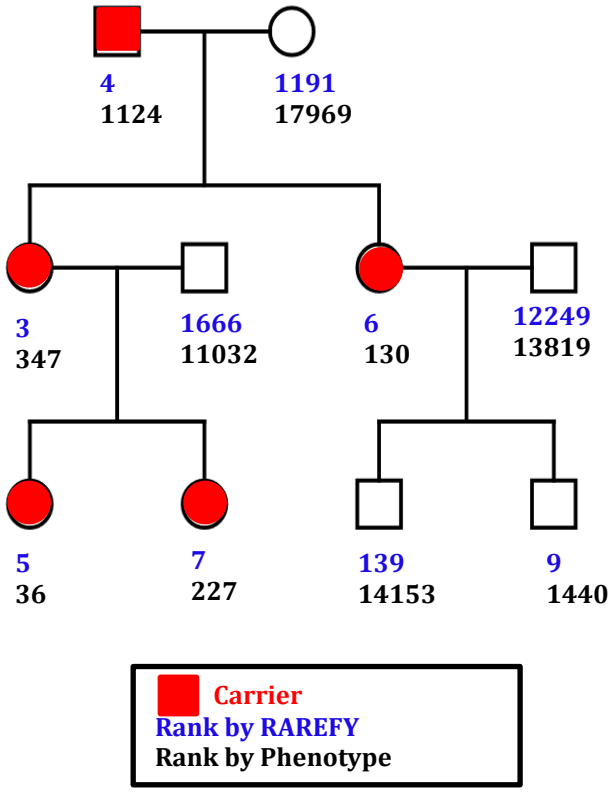


Figure S4. 1 An Example of Who Are Selected by RAREFY and Phenotypic Extremes Alone.

The example family was selected from a simulated sample with 20,000 individuals in 2,000 families. A trait-associated rare variant with frequency 0.001 and effect 2 standard deviation was simulated in founders and then gene-dropped to children. Blue numbers represent rank of calculated RAREFY score. Black numbers represent rank of phenotype. Individuals marked as red in the pedigrees are carries.

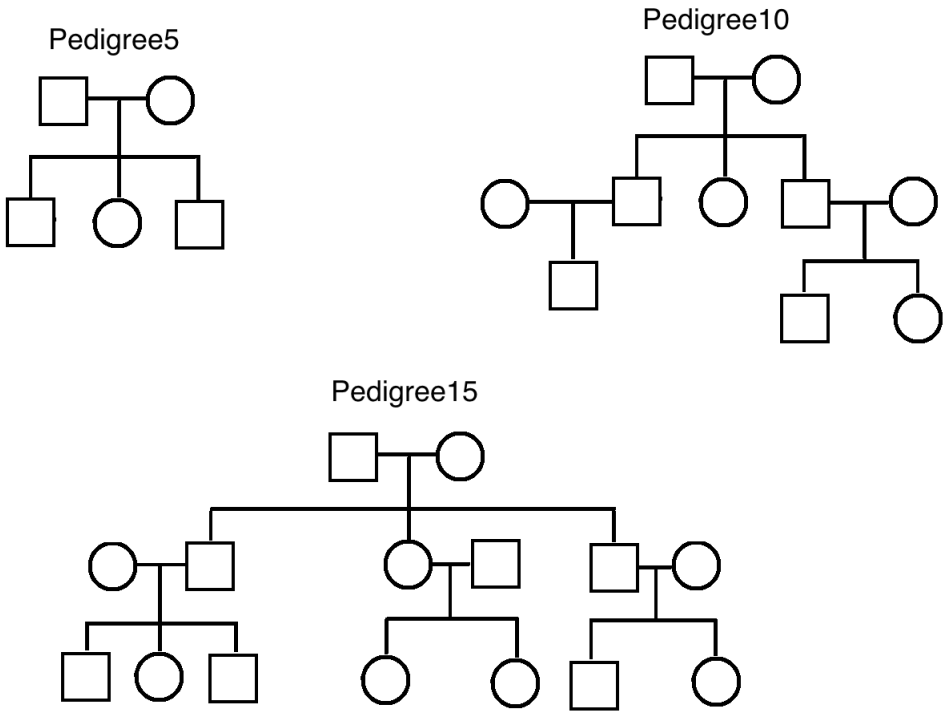


Figure S4. 2 Pedigree Structures Used in Simulations.

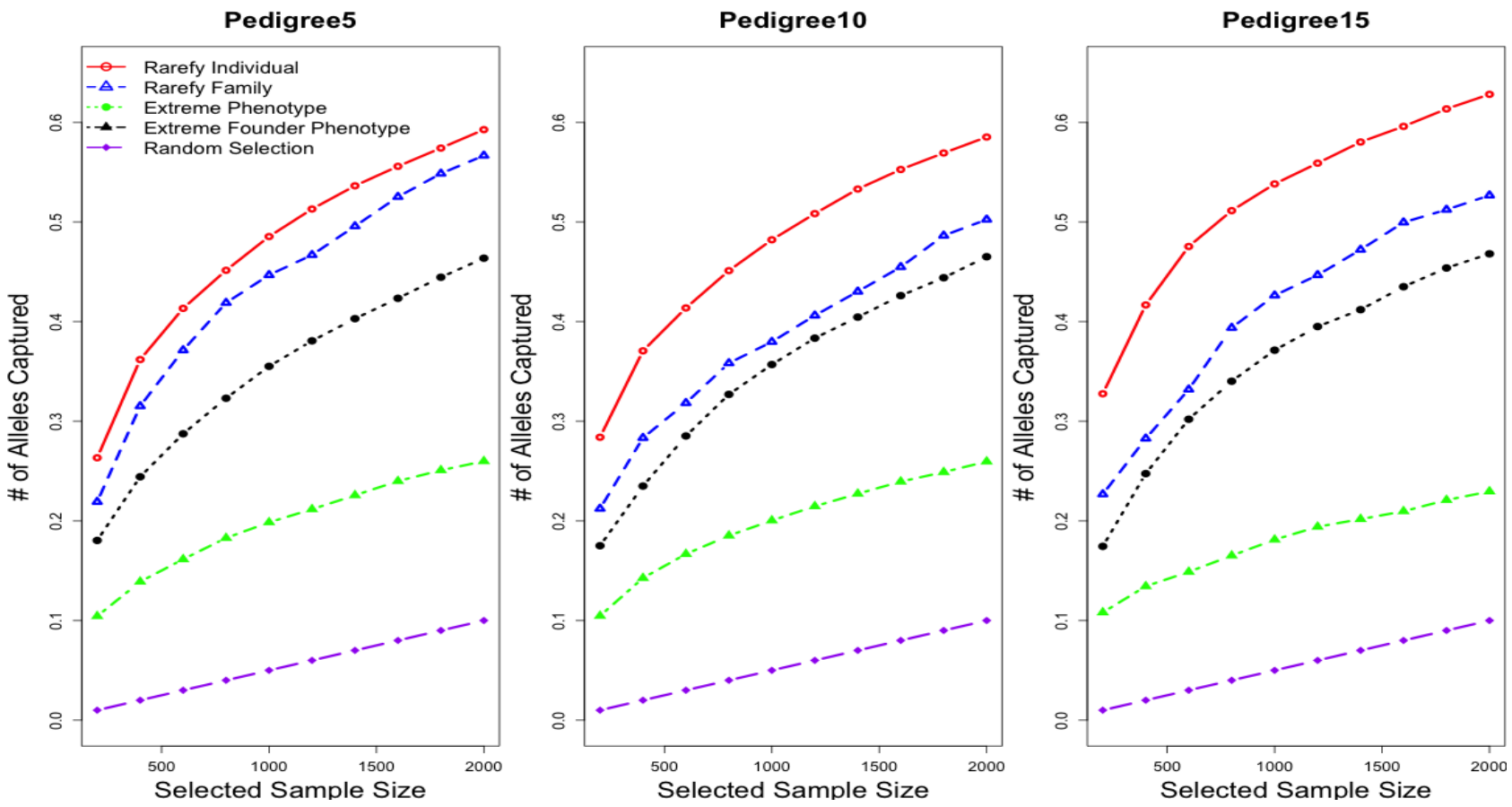


Figure S4. 3 Gene-level Discovery Power Selecting from 20,000 individuals and All Causal Variants were Trait-Increasing

DNA sequences of 1,000 base-pair were simulated in founders and then gene drop to children in various pedigree structures Pedigree5, Pedigree10, and Pedigree15 (See Figure S4.1). 20% of variants with frequency < 0.01 were selected as causal variants and explained 1% trait variance in total. All causal variants were trait-increasing. 200-2,000 individuals were selected from samples with 20,000 individuals.

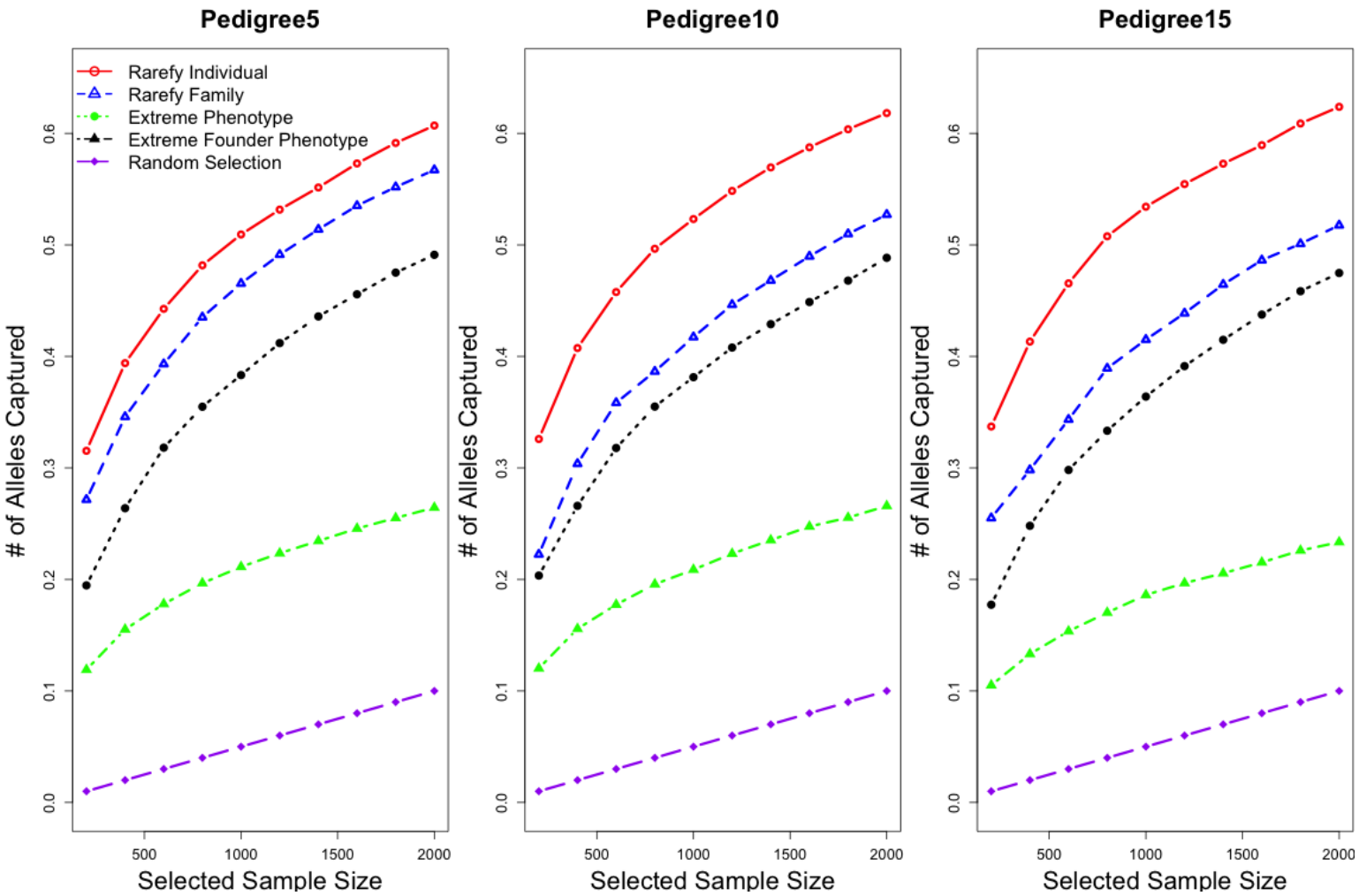


Figure S4. 4 Gene-level Discovery Power Selecting from 20,000 individuals and 50% Causal Variants were Trait-Increasing

DNA sequences of 1,000 base-pair were simulated in founders and then gene drop to children in various pedigree structures Pedigree5, Pedigree10, and Pedigree15 (See Figure S4.1). 20% of variants with frequency < 0.01 were selected as causal variants and explained 1% trait variance in total. All causal variants were trait-increasing. 200-2,000 individuals were selected from samples with 20,000 individuals.

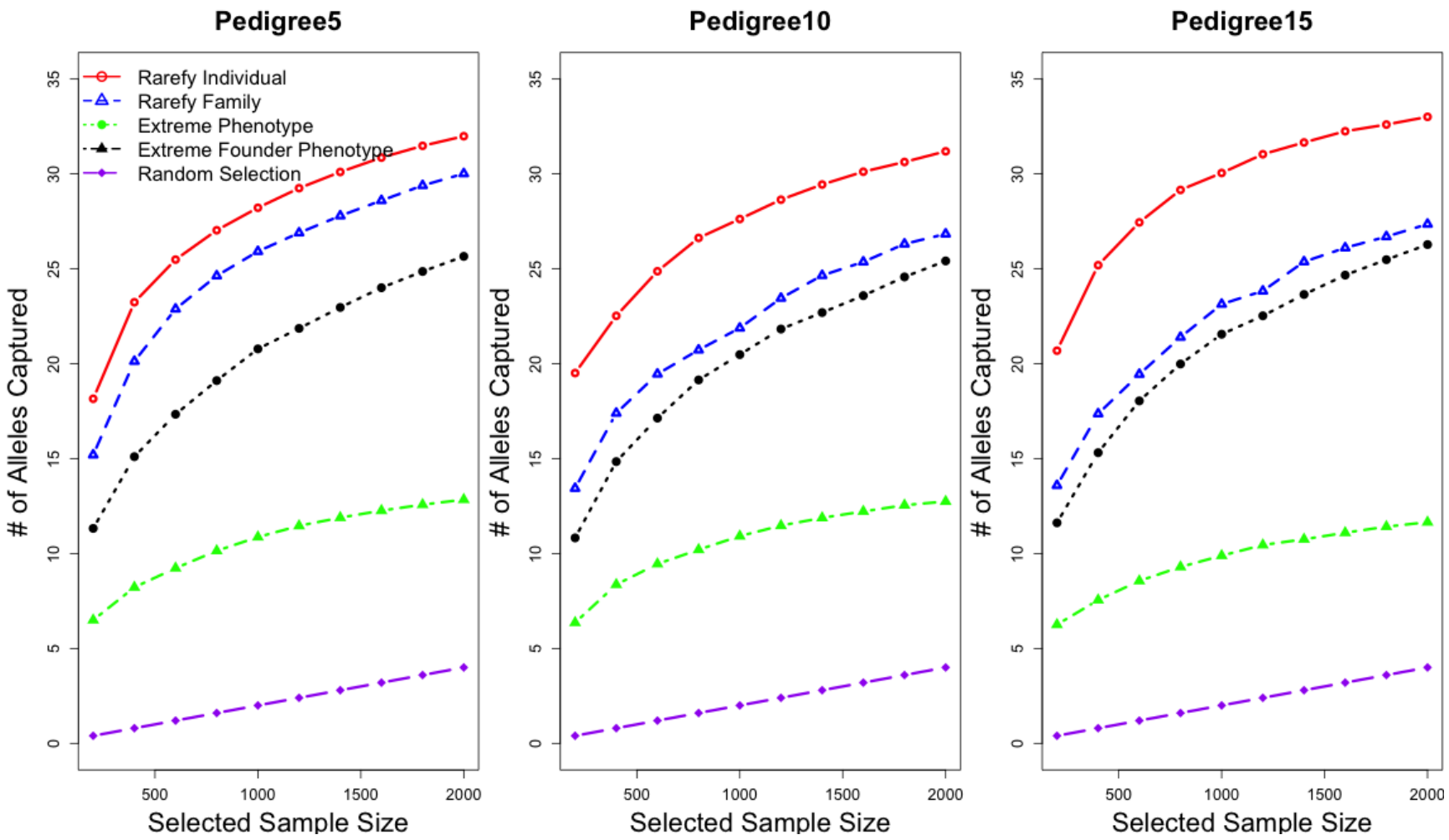


Figure S4. 5 Pedigree Structure and Size on Single Variant Discovery Power.

A single variant with frequency 0.001 and effect size 2 was simulated in founders and then gene-dropped to children in Pedigree5, Pedigree10, and Pedigree15. 200-2,000 individuals were selected from a sample with 20,000 individuals.

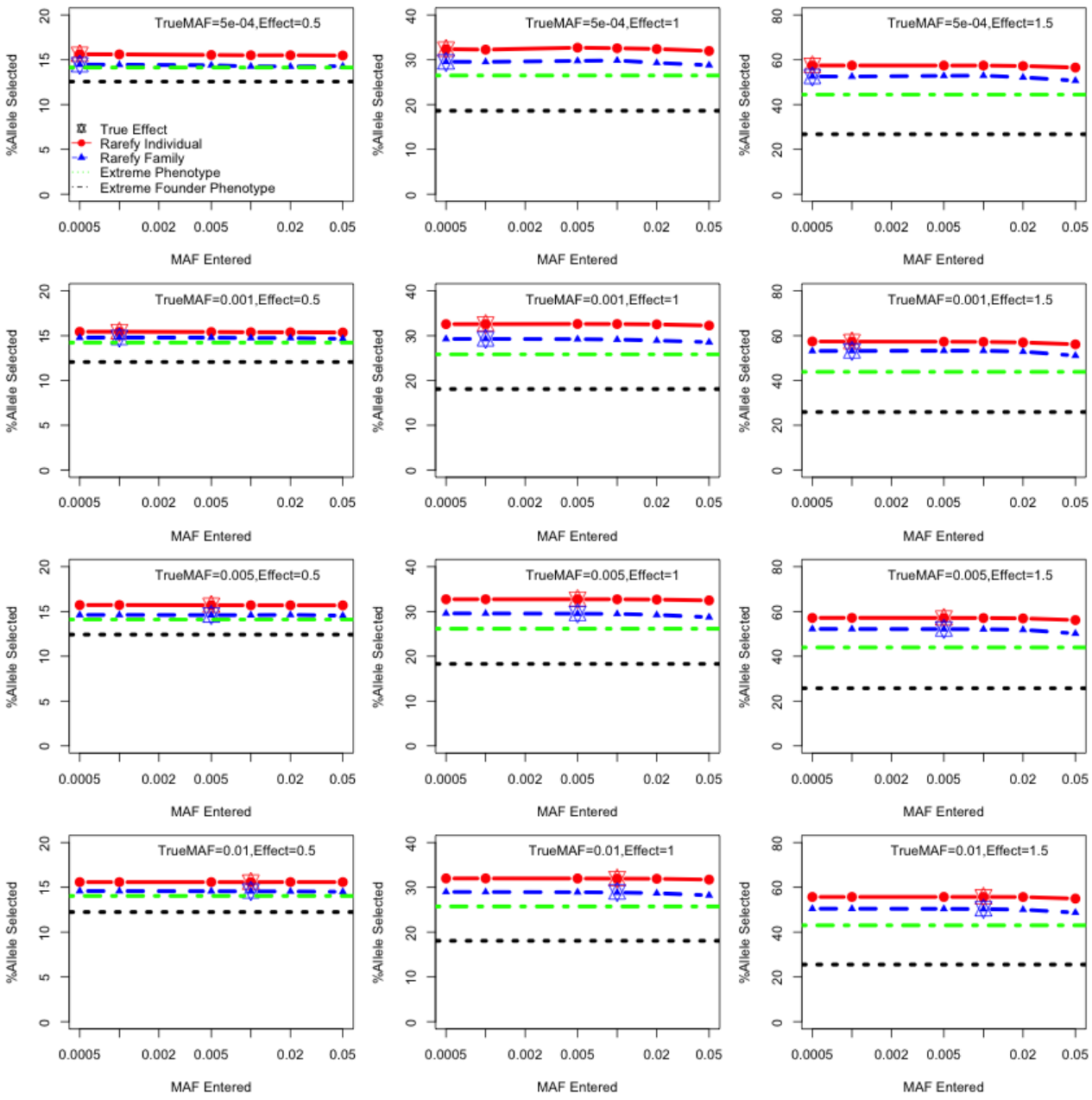


Figure S4. 6 Effect of Misspecifying MAF on Selection Efficiency

1000 family samples of 10,000 individuals with Pedigree5 (See **Figure1**) structure were used to collect each data point. True effect sizes were used in these simulations.

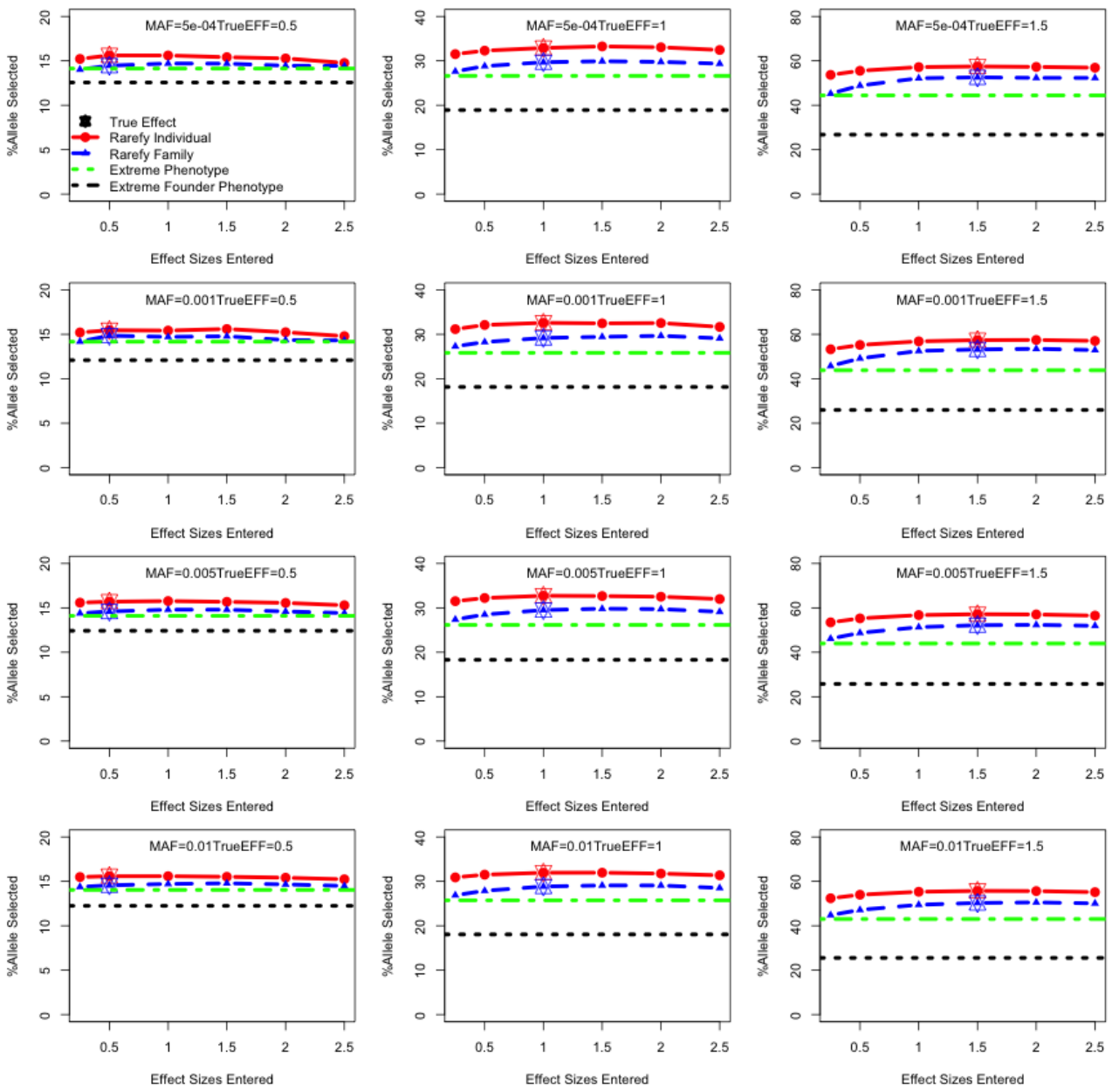


Figure S4. 7 Effect of Misspecifying Effect Sizes on Selection Efficiency

1000 family samples of 10,000 individuals with Pedigree5 (See **Figure1**) structure were used to collect each data point. True MAFs were used in these simulations.

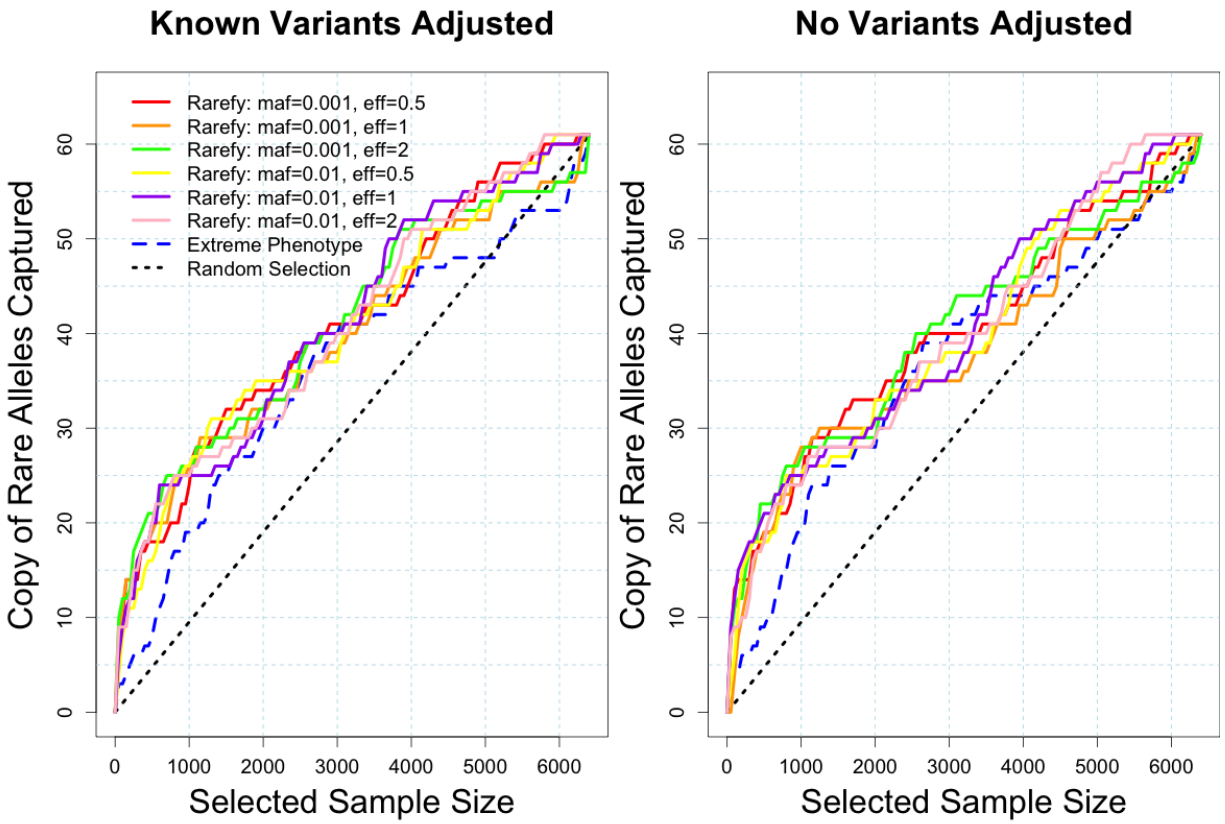


Figure S4. 8 Power to Discover Rare Allele of V578 from SardiNIA Sample Using RAREFY-Individual and Extreme Phenotype Strategies.

50-6,400 individuals were selected for sequencing from SardiNIA sample of 6410 phenotyped individuals and covariates adjusted. Adjusted LDL values (40 individuals taking cholesterol-lowering drugs were added 40 to their LDL values) were used as phenotype. Covariates were age, sex, and squared age, together with known variants (See **Table S4.1** for list of known variants adjusted).

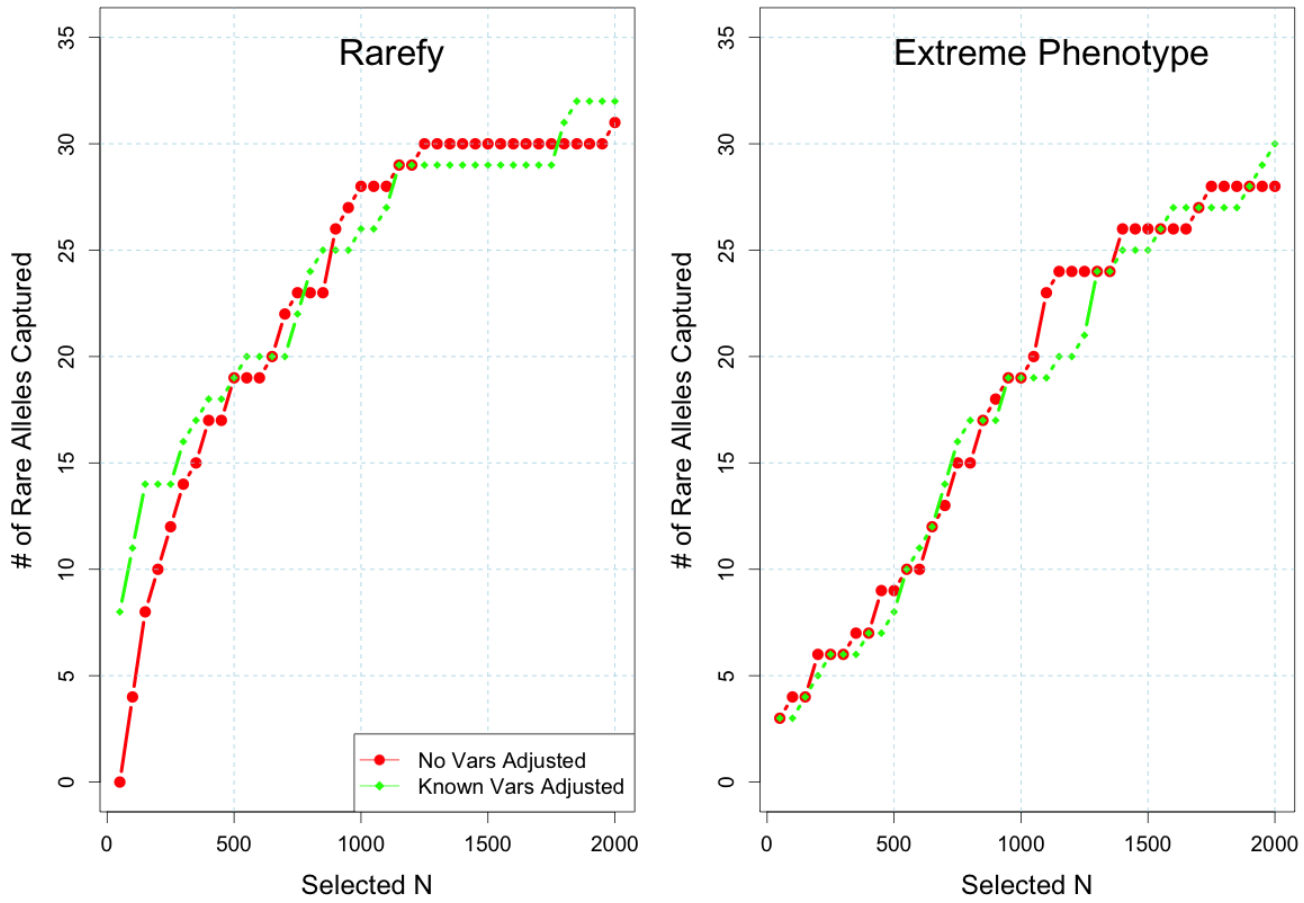


Figure S4. 9 Effect of Adjusting Known Variants to RAREFY and Extreme Phenotype Methods.

50-2,000 individuals were selected for sequencing from SardiNIA sample of 6410 phenotyped individuals and covariates adjusted. Adjusted LDL values (40 individuals taking cholesterol-lowering drugs were added 40 to their LDL values) were used as phenotype. Covariates were age, sex, and squared age, together with known variants (See **Table S4.1** for list of known variants adjusted). Default parameters (MAF=0.001 and effect size =1) were used running RAREFY.

Table S4. 1 Summary Statistics for LDL in SardiNIA Sample.

		Male						Female					
	Sample Size*	**N**	**Mean**	**Median**	**Min**	**Max**	**Age (mean, median)**	**N**	**Mean**	**Median**	**Min**	**Max**	**Age (mean, median)**
SardiNIA	6290	2670	129.5	128.2	27.4	330.5	44.1, 42.4	3620	125.9	132.4	27.9	293.3	43.7, 42.0

* Counting samples who were phenotyped and had age and sex recorded, and known variants genotyped (See **Table S4.2** for list of known variants adjusted).

Table S4. 2 Variants Known to be Associated to LDL and Adjusted in Analysis.

Chr:Pos	Status*	Chr:Pos	Status	Chr:Pos	Status	Chr:Pos	Status
1:25775733	included	2:27741237	included	6:161010118	included	12:121388962	included
1:55496039	included	2:44065090	included	7:21607352	included	12:121416650	included
1:55504650	included	2:44072576	included	7:44579180	included	14:24883887	included
1:55505647	included	2:44073881	included	8:9183358	included	16:56993324	included
1:63025942	included	4:89039082	included	8:9185146	included	16:72108093	included
1:63118196	included	5:74625487	included	8:59388565	included	17:45425115	included
1:109817590	included	5:74648603	included	8:126482077	included	19:11195030	not included
1:109817838	included	5:74651084	included	8:126490972	included	19:11202306	not included
1:109818306	included	5:74655726	included	8:126504726	included	19:11210912	not included
1:109818530	included	5:74656539	included	8:145043543	included	19:11238473	not included
1:109822166	included	5:156390297	included	9:91540059	included	19:19407718	not included
1:207875175	included	5:156398169	included	9:136155000	included	19:19407718	not included
1:220973563	included	6:16127407	included	10:113933886	included	19:19658472	not included
1:234858597	included	6:16161425	included	11:61569830	included	19:19789528	not included
2:20903015	included	6:16197194	included	11:61597212	included	19:22614122	not included
2:21231524	included	6:26093141	included	11:116603724	included	19:45395266	not included
2:21232195	included	6:32412435	included	11:116607437	included	19:45412079	not included
2:21232195	included	6:33143948	included	11:116648917	included	19:45422946	not included
2:21263900	included	6:116312893	included	11:116652423	included	20:39091487	included
2:21286057	included	6:131256364	included	11:126243952	included	20:39228784	included
2:21288321	included	6:160578860	included	12:112072424	included	20:39672618	included

*: Variants from chr19 were not included in analysis, in case of possible LD with LDLR V578 variant that we are interested in testing.

CHAPTER 5: SUMMARY AND FUTURE DIRECTIONS

Summary

In this dissertation, I discuss efficient study design for sequencing-based GWAS, describe a novel approach to prioritize informative families and individuals to sequence first, and proposed gene-level association methods for both single studies and meta-analyzing family and population samples across multiple studies for both autosomal and X-linked genes.

In chapter 2, I demonstrate that with the same sample size, family samples have larger power to detect trait-associated rare and low-frequency variants with moderate to large effect sizes because family samples can allow multiple copies of trait-associated rare alleles to be observed in a single family or a "Jackpot" effect. Although on average, when sample size is the same, rare allele count is expected to be the same in family samples and population samples, variability in allele counts is larger in families particularly in pedigrees with many descendants per founder. This power advantage is particularly obvious for variants with extremely low frequency, for example singletons and doubletons where population samples barely have any power, or low-frequency with large effect sizes. When variants become more frequent or effect sizes become moderate to low, population samples quickly catch up and exceed family samples in power.

In chapter2, I also describe a series of computationally efficient gene-level tests for family samples including burden, variable threshold, and sequence kernel association tests that are built upon single-variant summary statistics and covariance matrices without utilizing raw data. This also makes our methods to be easily adapted to gene-level tests in meta-analysis where raw data sharing is usually not feasible across studies. The fact that all gene-level statistics are reconstructed from single-variant scan makes it computationally efficient to perform multiple gene-level tests and using various grouping strategies without going through raw data repeatedly. Together with our efficient implementation in C++ code and computational considerations whenever possible in terms of speed and memory use, our software uses the least CPU time and memory compare to other implementations. Another major difference between our methods and others is that our variance component model is flexible enough to handle known familial relatedness, cryptic relatedness and population structure. Using simulations, I confirm that burden type tests are more powerful than kernel-based tests when number of causal variants in a gene is large and all causal variants have the same direction in effect sizes, and kernel-based tests are more powerful when number of causal variants in a gene is small or causal variants have opposite directions in effect sizes. I further demonstrate the usage of our method and software by meta-analyzing SardiNIA and HUNT exome chip data for HDL level by not only successfully detect known genes that are associated with HDL but also particularly detect *APOC3* with a smaller burden p-value using ~12,000 individuals than previous studies where 10 times more samples were analyzed [Crosby et al. 2014].

In chapter 3, I describe a series of gene-level tests for X-linked variants and genes in families and unrelated individuals. I describe the variance component model where autosome and chromosome X polygenic contributions are modeled separately. This variance component is flexible enough to model familial relatedness, cryptic relatedness, and population structure manifested through both autosome and chromosome X. I further describe the method to estimate relatedness from genotypes of X-linked markers to model phenotypic correlation contributed by X chromosome. Our work assumes complete X-inactivation in female, and code genotypes of females as 0,1, or 2 and males as 0 or 2 correspondingly. In this case, male samples contribute twice the variance of female samples toward total trait variance for genes of the same effect sizes. Using simulations, I show that our gene-level association methods and implementation are well controlled under the null. I also demonstrate that there is larger power to detect an X-linked gene than an autosomal gene of the same effect sizes, and more males in a sample provides larger power to detect an X-linked gene when sample size is fixed. Finally, I demonstrate the usage of our methods and tool using SardiNIA quantitative traits and report the association of *G6PD* with multiple traits measured in this study.

In chapter 4, I describe a novel likelihood-based approach to prioritize individuals who are likely to be carriers of traits associated rare variants, when budget is limited but sequencing cost is high. I describe the exact calculation for small pedigrees and an MCMC approach to estimate the quantity when exact calculation is not feasible to large pedigrees. By simulations, I demonstrate that our method has larger power in both capturing trait-associated rare alleles and detecting association of these variants compare

to selecting individuals with phenotypic extremes alone. I also demonstrate that the constant prior value of frequency and effect sizes have minimal effect on power if they are specified within a reasonable range. Using SardiNIA data where the largest family has ~1,200 individuals, I demonstrate the usage and computational cost of our method and implementation and show that RAREFY has larger discovery power than others.

Future Directions

There are many open questions in sequencing-based association studies. They can be statistical and computational challenges. In the future, I seek solutions or improvements in the following three topics.

First, I seek to develop more efficient statistical method or computational approach to fit variance component model in large samples. Our current method and implementation has been tested to be able to analyze 20,000 individuals and 200,000 variants in a week on *** CPUs. However, larger samples might be available in the near future, which requires more efficient implementation and statistical approach. Although a convenient way to analyze very large sample is to divide a sample into smaller ones that have the minimal relatedness in between, for example, a division by ancestry, the correlation between the sub samples is un-avoidable. Also, when multiple variance components are included in the model, for example, when chromosome X variance component or shared-environment component is included in the model, currently available fast algorithms for simple variance component model, which contains only one genetic component and non-shared

environment, are not applicable. Faster likelihood maximizing algorithms are in great need for this type of variance component models.

Second, there are many interesting topics that need further investigation in meta-analysis. In my dissertation, we propose methods to meta-analyze both related and unrelated samples or samples with population structure or cryptic relatedness. However, in real data analysis, relatedness between samples is possible. This type of relatedness, cryptic or distant, is usually ignored. The other topic in meta-analysis that might be interesting is that when meta-analyzing genotyping array data across studies, effect size estimates are not homogeneous even for the same variant because variants have very different linkage disequilibrium patterns in different populations. This issue has more impact on rare variants than on common ones because rare variant frequencies are more likely to be confounded with population structure and geological locations. Special considerations are needed for the existence of heterogeneity in meta-analysis.

Third, there are a few topics that are in need of discussion in chromosome X analysis. Our current approach has a few assumptions, for example, variants in male and female have the same frequency and effect sizes. These assumptions could easily be violated in real data. For example, a gene could be differently expressed between sexes in some tissues but not in others, which suggests that male and female might have different architecture of regulatory interactions. This could lead to different effect sizes between male and female. In this situation, sex and genotype interactions might be a reasonable way to evaluate association thoroughly. Also, when frequency for the same variant is

different between sexes, our basic assumption is violated. Then a different variance component model where both female and male variance components are included might be more appropriate.

Conclusion

New problems emerge with rapid improvement in technology in human genetic studies. With more biological insight of disease and human health-related traits are revealed, the need of appropriate statistical methods and efficient computational solutions will be in great need. To this end, I propose power statistical methods and provide computationally efficient tools to facilitate the science community with discoveries of rare and low-frequency variants that contribute to the majority of polymorphism from sequencing studies. I evaluate study designs and provide insight on powerful discoveries from sequencing-based association studies. I propose powerful approach to identify possible carriers of variants of interest to sequence first which could be easily imputed to the rest of the sample, to gain power with limited cost. These statistical approach and design strategies will facilitate scientific investigators for faster and powerful discoveries.

BIBLIOGRAPHY

Abecasis, G.R. (2002). MINX: MERLIN in X. http://csgsphumichedu/abecasis/Merlin/referencehtml.

Abecasis, G.R. (2011). ExomePick. http://genomesphumichedu/wiki/ExomePicks.

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

Abecasis, G.R., Cardon, L.R., and Cookson, W.O. (2000). A general test of association for quantitative traits in nuclear families. American journal of human genetics 66, 279-292.

Abecasis, G.R., Cherny, S.S., and Cardon, L.R. (2001a). The impact of genotyping error on family-based analysis of quantitative traits. European journal of human genetics : EJHG 9, 130-134.

Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics 30, 97-101.

Abecasis, G.R., Cookson, W.O., and Cardon, L.R. (2001b). The power to detect linkage disequilibrium with quantitative traits in selected samples. American journal of human genetics 68, 1463-1474.

Abifadel, M., Varret, M., Rabes, J.P., Allard, D., Ouguerram, K., Devillers, M., Cruaud, C., Benjannet, S., Wickham, L., Erlich, D., et al. (2003). Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nature genetics 34, 154-156.

Adams, A.M., and Hudson, R.R. (2004). Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics 168, 1699-1712.

Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S., et al. (2007). Medical sequencing at the extremes of human body mass. American journal of human genetics 80, 779-791.

Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. American journal of human genetics 62, 1198-1211.

Astle, W., and Balding, D.J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. statistical Science 24, 451-471.

Boucas, A.P., Brondani, L.A., Souza, B.M., Lemos, N.E., de Oliveira, F.S., Canani, L.H., and Crispim, D. (2013). The A allele of the rs1990760 polymorphism in the IFIH1 gene is associated with protection for arterial hypertension in type 1 diabetic patients and with expression of this gene in human mononuclear cells. PloS one 8, e83451.

Carrel, L., and Willard, H.F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434, 400-404.

Chen, H., Meigs, J.B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. Genetic epidemiology 37, 196-204.

Chen, W.M., and Abecasis, G.R. (2007). Family-based association tests for genomewide association scans. American journal of human genetics 81, 913-926.

Cheung, C.Y., Marchani Blue, E., and Wijsman, E.M. (2014). A statistical framework to guide sequencing choices in pedigrees. American journal of human genetics 94, 257-267.

Chow, J.C., Yen, Z., Ziesche, S.M., and Brown, C.J. (2005). Silencing of the mammalian X chromosome. Annual review of genomics and human genetics 6, 69-92.

Chung, R.H., Morris, R.W., Zhang, L., Li, Y.J., and Martin, E.R. (2007). X-APL: an improved family-based test of association in the presence of linkage for the X chromosome. American journal of human genetics 80, 59-68.

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

Clayton, D.G. (2009). Sex chromosomes and genetic association studies. Genome medicine 1, 110.

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305, 869-872.

Cohen, J.C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G.L., Grundy, S.M., and Hobbs, H.H. (2006). Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proceedings of the National Academy of Sciences of the United States of America 103, 1810-1815.

Crosby, J., Peloso, G.M., Auer, P.L., Crosslin, D.R., Stitziel, N.O., Lange, L.A., Lu, Y., Tang, Z.Z., Zhang, H., Hindy, G., et al. (2014). Loss-of-function mutations in APOC3, triglycerides, and coronary disease. The New England journal of medicine 371, 22-31.

Davies, R. (1980). The distribution of a linear combination of chi-square random variables. J R StatSoc Ser C Appl Stat 29.

De, G., Yip, W.K., Ionita-Laza, I., and Laird, N. (2013). Rare variant analysis for family-based design. PloS one 8, e48495.

Ekstrom, C.T. (2004). Multipoint linkage analysis of quantitative traits on sex-chromosomes. Genetic epidemiology 26, 218-230.

Falconer, D.S., and Mackay, T. (1996). Introduction to Quantitative Genetics.(Longman, Essex, England).

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

Feng, S., Pistis, G., Zhang, H., Zawistowski, M., Mulas, A., Zoledziewska, M., Holmen, O.L., Busonero, F., Sanna, S., Hveem, K., et al. (2015). Methods for association analysis and meta-analysis of rare variants in families. Genetic epidemiology 39, 227-238.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-220.

Fulker, D.W., Cherny, S.S., Sham, P.C., and Hewitt, J.K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. American journal of human genetics 64, 259-267.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. Journal of Computational and Graphical Statistics 1, 141–150.

Giorgio, P., Eleonora, P., Vrieze, S.I., Sidore, C., Steri, M., Danjou, F., Busonero, F., Mulas, A., Zoledziewska, M., Maschio, A., et al. (2014). Rare variants genotype imputation with thousands of study-specific whole-genome sequences: Implications for cost-effective study designs. European Journal of Human Genetics.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences of the United States of America 108, 11983-11988.

Guey, L.T., Kravic, J., Melander, O., Burtt, N.P., Laramie, J.M., Lyssenko, V., Jonsson, A., Lindholm, E., Tuomi, T., Isomaa, B., et al. (2011). Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genetic epidemiology.

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57, 97-109.

Hoffmann, T.J., Sakoda, L.C., Shen, L., Jorgenson, E., Habel, L.A., Liu, J., Kvale, M.N., Asgari, M.M., Banda, Y., Corley, D., et al. (2015). Imputation of the rare HOXB13 G84E mutation and cancer risk in a large population-based cohort. PLoS genetics 11, e1004930.

Holmen, O.L., Zhang, H., Fan, Y., Hovelson, D.H., Schmidt, E.M., Zhou, W., Guo, Y., Zhang, J., Langhammer, A., Lochen, M.L., et al. (2014a). Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. Nature genetics 46, 345-351.

Holmen, T.L., Bratberg, G., Krokstad, S., Langhammer, A., Hveem, K., Midthjell, K., Heggland, J., and Holmen, J. (2014b). Cohort profile of the Young-HUNT Study, Norway: A population-based study of adolescents. International journal of epidemiology 43, 536-544.

Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337-338.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., and Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. European journal of human genetics : EJHG.

J., B., and P., C. (1991). Sequential Monte Carlo p-values. Biometrika 78, 301-304.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature genetics 42, 348-354.

Kent, J.W., Jr., Dyer, T.D., and Blangero, J. (2005). Estimating the additive genetic effect of the X chromosome. Genetic epidemiology 29, 377-388.

Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proceedings of the National Academy of Sciences of the United States of America 106, 3871-3876.

Kwan, J.S., Cherny, S.S., Kung, A.W., and Sham, P.C. (2009). Novel sib pair selection strategy increases power in quantitative association analysis. Behavior genetics 39, 571-579.

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

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

Laird, N.M., and Lange, C. (2008). Family-based methods for linkage and association analysis. Advances in genetics 60, 219-252.

Lange, K. (1997). Mathematical and Statistical methods for genetic analysis.

Lange, K., and Sobel, E. (2006). Variance component models for X-linked QTLs. Genetic epidemiology 30, 380-383.

Lange, L.A., Hu, Y., Zhang, H., Xue, C., Schmidt, E.M., Tang, Z.Z., Bizon, C., Lange, E.M., Smith, J.D., Turner, E.H., et al. (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. American journal of human genetics 94, 233-245.

Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., and Lin, X. (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. American journal of human genetics 91, 224-237.

Lee, S., Teslovich, T.M., Boehnke, M., and Lin, X. (2013). General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies. American journal of human genetics 93(1), 42-53.

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

Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American journal of human genetics 83, 311-321.

Lin, D.Y., and Sullivan, P.F. (2009). Meta-analysis of genome-wide association studies with overlapping subjects. American journal of human genetics 85, 862-872.

Lin, D.Y., and Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. American journal of human genetics 89, 354-367.

Lin, D.Y., and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika 97, 321-332.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nature methods 8, 833-835.

Liu, D.J., Peloso, G.M., Zhan, X., Holmen, O.L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P.L., Goel, A., Zhang, H., et al. (2014). Meta-analysis of gene-level tests for rare variant association. Nature genetics 46, 200-204.

Liu, H., Tang, Y., and Zhang, H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Comput Stat Data Anal 53, 853-856.

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

Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., et al. (2011). The functional spectrum of low-frequency coding variation. Genome biology 12, R84.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. J Chem Phys 21.

Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology 34, 188-193.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337, 100-104.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature 456, 98-101.

Ober, C., Loisel, D.A., and Gilad, Y. (2008). Sex-specific genetic architecture of human disease. Nature reviews Genetics 9, 911-922.

Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. Nature reviews Genetics 12, 465-474.

Payer, B., and Lee, J.T. (2008). X chromosome dosage compensation: how mammals keep the balance. Annual review of genetics 42, 733-772.

Pilia, G., Chen, W.M., Scuteri, A., Orru, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS genetics 2, e132.

Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. American journal of human genetics 86, 832-838.

Raychaudhuri, S., Iartchouk, O., Chin, K., Tan, P.L., Tai, A.K., Ripke, S., Gowrisankar, S., Vemuri, S., Montgomery, K., Yu, Y., et al. (2011). A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. Nature genetics 43, 1232-1236.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science 273, 1516-1517.

Risch, N., and Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268, 1584-1589.

Risch, N.J., and Zhang, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. American journal of human genetics 58, 836-843.

Saad, M., and Wijsman, E.M. (2014). Power of family-based association designs to detect rare variants in large pedigrees using imputed genotypes. Genetic epidemiology 38, 1-9.

Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H.M., Jackson, A.U., Piras, M.G., Usala, G., Maninchedda, G., Sassu, A., et al. (2011). Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS genetics 7, e1002198.

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

Schifano, E.D., Epstein, M.P., Bielak, L.F., Jhun, M.A., Kardia, S.L., Peyser, P.A., and Lin, X. (2012). SNP Set Association Analysis for Familial Data. Genetic epidemiology.

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., et al. (2007). A genome-wide association

study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316, 1341-1345.

Staples, J., Nickerson, D.A., and Below, J.E. (2013). Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. Genetic epidemiology 37, 136-141.

Svishcheva, G.R., Belonogova, N.M., and Axenovich, T.I. (2014). FFBSKAT: fast family-based sequence kernel association test. PloS one 9, e99407.

T2D-GENES-Consortium. (,In Preparation). Evaluating the Rare Variant Contribution to Complex Disease in Pedigrees. In Preparation.

Tang, Z.Z., and Lin, D.Y. (2013). MASS: meta-analysis of score statistics for sequencing studies. Bioinformatics.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64-69.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-713.

Thornton, T., Zhang, Q., Cai, X., Ober, C., and McPeek, M.S. (2012). XM: association testing on the X-chromosome in case-control samples with related individuals. Genetic epidemiology 36, 438-450.

Van Gestel, S., Houwing-Duistermaat, J.J., Adolfsson, R., van Duijn, C.M., and Van Broeckhoven, C. (2000). Power of selective genotyping in genetic association analyses of quantitative traits. Behavior genetics 30, 141-146.

Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190-2191.

Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M., et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nature genetics 40, 161-169.

Wise, A.L., Gyi, L., and Manolio, T.A. (2013). eXclusion: toward integrating the X chromosome in genome-wide association analyses. American journal of human genetics 92, 643-647.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics 89, 82-93.

Yan, Q., Tiwari, H.K., Yi, N., Lin, W.Y., Gao, G., Lou, X.Y., Cui, X., and Liu, N. (2014). Kernel-machine testing coupled with a rank-truncation method for genetic pathway analysis. Genetic epidemiology 38, 447-456.

Zhan, X., Larson, D.E., Wang, C., Koboldt, D.C., Sergeev, Y.V., Fulton, R.S., Fulton, L.L., Fronick, C.C., Branham, K.E., Bragg-Gresham, J., et al. (2013). Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. Nature genetics 45, 1375-1379.

Zhang, L., Martin, E.R., Chung, R.H., Li, Y.J., and Morris, R.W. (2008). X-LRT: a likelihood approach to estimate genetic risks and test association with X-linked markers using a case-parents design. Genetic epidemiology 32, 370-380.

Zhang, L., Martin, E.R., Morris, R.W., and Li, Y.J. (2009). Association test for X-linked QTL in family-based designs. American journal of human genetics 84, 431-444.

Zheng, G., Joo, J., Zhang, C., and Geller, N.L. (2007). Testing association for markers on the X chromosome. Genetic epidemiology 31, 834-843.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nature genetics 44, 821-824.