ASSOCIATION ANALYSIS OF RARE VARIANTS IN SEQUENCING STUDIES

Zhengzheng Tang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2014

Approved by:

Dr. Danyu Lin

Dr. Donglin Zeng

Dr. Wei Sun

Dr. Yun Li

Dr. Matthew R. Nelson

ABSTRACT

ZHENGZHENG TANG: Association Analysis of Rare Variants in Sequencing Studies
(Under the direction of Dr. Danyu Lin)

Recent advances in sequencing technologies have made it possible to explore the influence of rare variants on complex diseases and traits. Large-scale sequencing studies provide the opportunity to examine the proportion of the missing heritability that is attributable to rare variants. They also pose a range of analytical and computational challenges that cannot be adequately addressed with existing methods.

For the association analysis of the rare variants, it is customary to aggregate rare mutations within a gene to perform gene-level association analysis. In the first part of the dissertation, we develop asymptotic and resampling gene-level association tests for a variety of traits and study designs. We employ score statistics under appropriate statistical models to achieve numerical stability and computational efficiency. The resulting software SCORE-Seq features a large collection of utilities devoted to perform gene-level association analysis in different scenarios.

Trait-dependent sampling has been adopted in many sequencing projects to reduce cost. In the second part, we provide a valid and efficient maximum likelihood framework for analyzing binary secondary traits under such sampling strategy. We produce the commonly used gene-level association tests and compare our methods with the naïve methods ignoring the trait-dependent sampling.

A single sequencing study is often underpowered to detect modest genetic effect of rare variants. Several methods are available to conduct meta-analysis for rare variants under fixed-effects models, which assume that the genetic effects are the same across all studies. In practice, genetic associations are likely to be heterogeneous among studies because of differences in population composition, environmental factors, phenotype and genotype measurements, or analysis method. In the third part, we propose a general framework for meta-analysis of sequencing studies that allows the genetic effects to vary among studies. We produce the fixed-effects and random-effects versions of all commonly used gene-level association tests. Our methods take score statistics, rather than individual participant data, as input and thus can accommodate any study designs and any phenotypes. We demonstrate through extensive simulation studies that our tests are more powerful than the existing ones in a wide range of practical situations.

I dedicate this dissertation work to my parents,
Jianhua Tang and Ziping Luo,
who have loved and supported me throughout my life,
and to my beloved husband and son,
Guanhua Chen and Patrick L. Chen
who stood by me through the good times and bad.

ACKNOWLEDGMENTS

My graduate experience at University of North Carolina at Chapel Hill has been an amazing journey. I am grateful to a number of people who have guided and supported me throughout the research process, and cheered me during my venture.

My deepest gratitude is to my advisor, Dr. Danyu Lin, for his guidance, support and patience. I have been very fortunate to have an advisor like him. And I would not have been able to achieve this accomplishment without him.

I would also like to thank my committee members Dr. Donglin Zeng, Dr. Wei Sun, Dr. Yun Li and Dr. Matthew R. Nelson for their insightful comments and constructive criticisms at different stages of my research. These comments motivated many of my thinking.

I am also thankful to Dr. Lloyd E. Chambless who supported me in my early years in CSCC. The experience of collaboration under his guidance was invaluable.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES xi

1 Literature review 1

1.1 Introduction 1

1.2 Gene-level association tests 3

1.2.1 Burden tests: CAST, GRANVIL, CMC 5

1.2.2 Weighted approach: WSS, KBAC 6

1.2.3 Maximization approach: VT 6

1.2.4 Signed approach: Han and Pan 7

1.2.5 Variance-Component (VC) tests: Similarity regression, C-alpha, SKAT 7

1.3 Trait-dependent sampling 8

1.3.1 Estimating parameters for the primary trait 10

1.3.2 Estimating parameters for the secondary trait 12

1.3.3 Performing association tests 12

1.4 Meta-analysis 14

1.4.1 Meta-analysis of GWAS 14

1.4.2 Meta-analysis of sequencing studies 17

2 A general framework for detecting disease associations with rare variants in sequencing studies 21

2.1 Introduction 21

2.2 Methods 22

2.2.1 Binary phenotypes 23

2.2.2 Quantitative phenotypes 29

2.2.3 Survival outcomes 30

2.2.4 Family-based data 32

2.3 Simulation studies 34

2.4 Data analysis 47

2.5 Discussion 58

3 Binary secondary trait analysis under trait-dependent sampling 61

3.1 Introduction 61

3.2 Methods 61

3.3 Simulation studies 64

3.3.1 Type I error under random sampling 64

3.3.2 Type I error and power under trait-dependent sampling 67

3.3.3 Meta-analysis 68

3.4 Data analysis 72

4 Meta-analysis in sequencing studies 77

4.1 Introduction 77

4.2 Methods 78

4.3 Simulation studies 83

4.4 Data analysis 90

4.5 Discussion 92

Appendix : Test statistics in Chapter 4 99

BIBLIOGRAPHY 102

LIST OF TABLES

2.1 Type I error^{a}and power of asymptotic methods with different weight functions 38

2.2 Type I error^{a}and power of asymptotic and permutation methods 38

2.3 Type I error^{a}and power of fixed and variable threshold methods 39

2.4 Type I error^{a}and power of asymptotic and permutation tests for detecting potentially opposite effects 40

2.5 Type I error^{a}and power of fixed and variable threshold methods with covariates 41

2.6 Type I error^{a}and power of asymptotic and bootstrap tests for detecting potentially opposite effects in the presence of covariates 42

2.7 Type I error^{a}and power of asymptotic methods with different weight functions under $p_j = 0.0005j$ $(j = 1, \cdots, 10)$ 43

2.8 Type I error^{a}and power of asymptotic methods with different weight functions under $p_j = 0.00025j$ $(j = 1 \cdots, 20)$ 43

2.9 Type I error^{a}and power of asymptotic methods with different weight functions under $p_j = 0.005$ $(j = 1, \cdots, 10)$ 44

2.10 Type I error^{a}and power of asymptotic methods with different weight functions under $p_j = 0.0025$ $(j = 1, \cdots, 10)$ 44

2.11 Type I error^{a}and power of asymptotic methods with different weight functions under $p_j = 0.0025$ $(j = 1, \cdots, 20)$ 45

2.12 Type I error^{a}of score, Wald and LR tests with covariates 45

2.13 Power of score, Wald and LR tests with covariates under $H_1 : \beta_j = x$ $(j = 1, \ldots, 10)$ 46

2.14 Power of score, Wald and LR tests with covariates under $H_1 : \beta_j = x/\{p_j(1 - p_j)\}^{1/2}$ $(j = 1, \ldots, 10)$ 46

3.1 Type I error^{a}at different nominal levels for rare-variant tests based on probit and logistic regressions under $p_j = 0.001j$ $(j = 1, \cdots, 10)$. Binary data are simulated using a logistic regression model. 65

3.2 Type I error^{a}at different nominal levels for rare-variant tests based on probit and logistic regressions under $p_j = 0.01j$ $(j = 1, \cdots, 10)$. Binary data are simulated using a logistic regression model. 66

4.1 Type I error divided by the nominal significance level α for various meta-analysis methods 86

LIST OF FIGURES

2.1 Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the asymptotic T1, T5, MB_p and VT tests in the quantitative trait analysis of total cholesterol. 50

2.2 Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the permutation EREC, T5, MB_p and VT tests in the quantitative trait analysis of total cholesterol. 51

2.3 Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the asymptotic T1, T5, MB_p and VT tests in the binary trait analysis of total cholesterol. 52

2.4 Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the bootstrap EREC, T5, MB_p and VT tests in the binary trait analysis of total cholesterol. 53

2.5 Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the SKAT in the quantitative and binary trait analyses of total cholesterol (with covariates) and for the SKAT and C-alpha in the binary trait analysis of total cholesterol without covariates. 54

2.6 Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the asymptotic T1, T5, MB_p and VT tests in the binary trait analysis of total cholesterol without covariates. 55

2.7 Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the permutation EREC, T5, MB_p and VT tests in the binary trait analysis of total cholesterol without covariates. 56

2.8 Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the Han-Pan test, Price et al.'s VT test, and the asymptotic and permutation versions of the MB test in the binary trait analysis of total c-holesterol without covariates. 57

3.1 Type I error rates for the ML and naïve methods for detecting genetic associations when the genetic effects are positive. The results for the burden, VT, and VC tests are shown in the upper, middle, and lower rows, respectively. The left panel shows the power as a function of the genetic effect on the primary trait when the correlation between the primary and secondary traits is 0.4. The right panel shows the power as a function of the correlation between the primary and secondary traits when the genetic effect on the primary trait is 0.4. 69

3.2 Power of the ML and naïve methods for detecting genetic associations when the genetic effects are positive. The results for burden, VT, and VC tests are shown in the upper, middle, and lower rows, respectively. The left panel shows the power as a function of the genetic effect on the primary trait when the correlation between the primary and secondary traits is 0.4. The right panel shows the power as a function of the correlation between the primary and secondary traits when the genetic effect on the primary trait is 0.4. 70

3.3 Power of the ML and naïve methods for detecting genetic associations when the genetic effects are negative. The results for burden, VT, and VC tests are shown in the upper, middle, and lower rows, respectively. The left panel shows the power as a function of the genetic effect on the primary trait when the correlation between the primary and secondary traits is 0.4. The right panel shows the power as a function of the correlation between the primary and secondary traits when the genetic effect on the primary trait is 0.4. 71

3.4 Power of the meta-analysis using the ML and naïve methods for detecting genetic associations. The results for the burden, VT, and VC tests are shown in the upper, middle, and lower rows, respectively. For the left panel, we set the genetic effect on the primary trait in both studies to 0.4 and the correlations between the primary trait and the secondary trait in Study I and Study II to 0.6 and -0.6, respectively. For the right panel, we set the correlations between the primary trait and the secondary trait in both studies to 0.4 and the genetic effect on the primary trait in Study I and Study II to 0.6 and -0.6, respectively. 73

3.5 Quantile-quantile plots for the ML, naïve-M', and naïve-C' meta-analysis results. 75

3.6 Pair plots between the p-values for the ML method and the naïve methods. 76

4.1 Power as a function of the between-study heterogeneity for the quantitative trait. The left, middle and right panels correspond to three different genetic structures: (a) genetic effects exhibit at the burden score level for variants with MAFs< 5%, (b) genetic effects exhibit at the burden score level for variants with MAFs< 1%, and (c) genetic effects exhibit at the individual variant level. For each structure, 80%, 50% or 20% of the variants in ten 300 base-pair regions were randomly selected to be potentially causal. 87

4.2 Power as a function of the between-study heterogeneity for the binary trait. The left, middle and right panels correspond to three different genetic structures: (a) genetic effects exhibit at the burden score level for variants with MAFs< 5%, (b) genetic effects exhibit at the burden score level for variants with MAFs< 1%, and (c) genetic effects exhibit at the individual variant level. For each structure, 80%, 50% or 20% of the variants in ten 300 base-pair regions were randomly selected to be potentially causal. 88

4.3 Power as a function of the number of studies with genetic effects on the quantitative trait. The upper, middle and lower panels correspond to three different genetic structures: (a) genetic effects exhibit at the burden score level for variants with MAFs< 5%, (b) genetic effects exhibit at the burden score level for variants with MAFs< 1%, and (c) genetic effects exhibit at the individual variant level. For each structure, 50% of the variants in ten 300 base-pair regions were randomly selected to be potentially causal. 89

4.4 Meta-analysis of the eleven studies in the NHLBI ESP: the left and middle panels are the quantile-quantile plots for the FE and RE tests, and the right panel compares the RE and FE results. The red dot indicates the gene LDLR. 93

4.5 Forest plots for the burden tests with two MAF thresholds for the gene LDLR in the NHLBI E-SP. For each study, the estimate of the genetic effect is shown by the square and the corresponding 95% confidence interval is shown by the line. The meta-estimate of the genetic effect and the corresponding 95% confidence interval are shown by the diamond. 94

4.6 Standardized LDL values for the carriers of the LDLR mutations in the NHLBI ESP. Each point represents an individual who carries a mutation. There are 54 polymorphic sites with the chr:position IDs and MAFs labeled on the x-axis. The variants are ordered by the MAFs. The vertical line pertains to the MAF threshold at which the test statistics of FE-VT and RE-VT are maximized. The five phenotype groups are indicated by different colors. AA and EA subjects are shown in circles and triangles, respectively. The horizontal line pertains to the average LDL value among individuals who do not carry any mutation. 95

CHAPTER1: LITERATURE REVIEW

1.1 Introduction

Complex diseases, such as cancer, hypertension, and diabetes, are determined by a variety of genetic and environmental factors, as well as their interactions. Genetic dissection of complex human disease is typically accomplished using data from large-scale genetic association studies, which explore relationship between genetic variants and disease phenotypes. Biological and empirical evidence suggests that rare variants account for a significant proportion of the genetic contribution to complex human diseases. Recent technological advances in next-generation sequencing (NGS) platforms have made it possible to generate comprehensive information on rare variants in large samples. Indeed, the number of sequencing studies has been increasing dramatically due to the widespread availability of NGS technologies and decrease in costs.

Gene-level testing is widely used in rare-variant association studies; however, the analytical methods must be tailored for different outcomes and study designs. In addition, the use of asymptotic approximations to assess the statistical significance has notable limitations in the setting of rare-variant testing. For instance, due to the low frequency of rare variants, asymptotic approximation may be violated, which can lead to inflated type I error and loss of power. Furthermore, the analytical distribution of the test statistic may not be known, and so, the statistical significance has to be evaluated empirically. Valid resampling methods for gene-level tests would therefore be desirable.

Although next-generation sequencing is much more cost-effective than Sanger sequencing, it is not economically feasible to sequence all study subjects in a very large cohort. A cost-effective strategy is to sequence only those subjects with extreme values of a quantitative trait. In the National Heart, Lung, and Blood Institute Exome Sequencing Project (NHLBI ESP), subjects with the highest or lowest values for body mass index (BMI), low-density lipoprotein (LDL), or blood pressure (BP) were selected for whole-exome sequencing. Failure to account for such trait-dependent sampling can cause severe inflation of type I error and substantial loss of power in quantitative trait analysis, especially when combining results from multiple studies that used different selection criteria. Thus, the valid and efficient statistical methods are needed for rare-variant association testing under such sampling design.

Due to the limited carriers of the rare mutations and high background rates of neutral variation even in causal genes, a single study is often underpowered to identify rare variants. Thus, meta-analysis becomes an important tool to increase statistical power by combining summary statistics over multiple studies. Fixed-effects models have been adopted almost exclusively for meta-analysis in genetic research. However, fixed-effects methods lose power if gene-level associations are heterogeneous among studies.

In this proposal, we first conduct a literature review in Chapter 1. In Chapter 2, we introduce a general framework for gene-level association analysis and develop asymptotic and resampling methods for different traits and study designs. In Chapter 3, we investigate a maximum likelihood framework to analyze binary secondary traits under trait-dependent sampling. Finally, in Chapter 4, we propose methods for meta-analysis of sequencing studies.

1.2 Gene-level association tests

Genome-wide association studies (GWAS) using tagSNPs (representative single nucleotide polymorphism in a region of the genome) have successfully identified common SNPs with small to modest effects for virtually every complex human disease. The standard approach for analyzing GWAS data is to apply a univariate test at each variant and then assess significance by using an appropriate p-value threshold, taking into account multiple testing. For dichotomous disease traits, commonly used association tests include χ^2 test, Fisher's exact test, alleles test, Armitage trend test, and tests based on logistic regression. The χ^2 test and Fisher's exact test can be employed for testing recessive, dominant, or codominant modes of inheritance. The alleles test and the Armitage trend test are tests of the additive mode of inheritance, in which the genotypes 0/1/2 are viewed as ordered categories. The regression approach is more flexible because it allows for covariate adjustment (e.g., principle components, environmental factors, and interactions) and different types of phenotypes (e.g., dichotomous disease indicator, count, or continuous measurement).

Genetic association studies may test hundreds of thousands of genetic variants for association with disease. Failure to account for the effects of multiple comparisons may result in an abundance of false positive results. Several approaches have been developed to correct for multiple testing. The simplest approach is to use a Bonferroni correction. In a typical GWAS study, the p-value cutoff for declaring significance is 5×10^{-8}. This figure is based on the approximate number of independent common variants across the genome. However, the Bonferroni correction is highly conservative, especially when the variants are in strong linkage disequilibrium. Permutation and Monte-Carlo methods (Lin 2005) are common alternative approaches to control false positive rates.

Multivariate methods can be used to combine information across variants and test for disease association with multiple variants simultaneously. Such tests not only capture the linkage disequilibrium patterns, but also decrease the number of required tests and thus reduce the penalty for multiple testing. The multivariate approach may be more powerful than the single-variant test if the variants have moderate effect sizes. Commonly used multivariate tests include Hotelling's T^2 test and multivariate tests based on regression. However, the large degrees of freedom compromise the power of these tests. In addition, simulation shows that Hotelling's T^2 test is highly sensitive to allele frequencies and power reduces drastically when the number of variants increase (Li and Leal 2008).

Technological advances in NGS platforms have made it possible to extend association studies to rare variants in targeted exons and eventually, the entire genome. Rare variants are believed to be enriched for functional alleles and have stronger effects on complex diseases than common variants (Pritchard 2001, Gorlov et al. 2008). Indeed, deep-resequencing studies of candidate genes have already demonstrated the influence of rare variants on several complex traits (Cohen et al. 2004, Ahituv et al. 2007, Nejentsev et al. 2009). Single-variant analysis has limited power in rare-variant association studies because only a small percentage of study subjects carry a rare mutation and adjustments need to be made for multiple testing. Other methods have been developed for detecting rare-variant associations (Tzeng and Zhang 2007, Li and Leal 2008, Madsen and Browning 2009, Han and Pan 2010, Liu and Leal 2010, Price et al. 2010, Wu et al. 2011, Sun et al. 2013). These methods are usually called "gene-level" methods because they combine information across multiple variant sites within a gene and the tests are performed for individual genes instead of individual variants. The gene-level methods can enrich association signals and reduce the penalty for multiple testing. In addition, prior biological knowledge (e.g., variant function, deleterious prediction) can be used to refine the grouping of variants. In the following sections, we provide extensive reviews of different gene-level methods.

1.2.1 Burden tests: CAST, GRANVIL, CMC

The burden tests generate genetic variable(s) by collapsing variants on the basis of specific criteria and applying univariate or multivariate tests for analysis of the genetic variable(s). The commonly-used criterion is to aggregate variants with minor allele frequencies (MAFs) less than a certain frequency threshold, and most burden tests are based on one genetic variable. For example, the Cohort Allelic Sums Test (CAST) collapses all variants below some frequency threshold and contrasts the number of individuals with one or more mutations between cases and controls (Morgenthaler and Thilly 2007). The Gene- or Region-based ANanlysis of Variants of Intermediate and Low frequency test (GRANVIL) is another burden test similar to CAST, in which the likelihood ratio test is performed under a linear regression framework (Morris and Zeggini 2010).

The CAST and GRANVIL tests enrich the association signals and reduce the number of degrees of freedom for testing. However, the inclusion of non-causal variants or the exclusion of causal variants during collapsing dilutes the association signal and adversely affects the power. The Combined Multivariate and Collapsing (CMC) test was developed to harness the advantages of both the collapsing and multivariate tests (Li and Leal 2008). For the CMC test, variants are divided into rare and common groups based on an allele frequency cutoff. In particular, rare variants (e.g., those with MAFs < 0.01) are collapsed together, whereas each common variant forms a separate group. Within each group, the individuals are coded as 1 if they carry one or more mutations and coded as 0 otherwise. A multivariate test (e.g., Hotelling's T^2 test) is then applied for detecting diseases associated with those genetic variables. The CMC test is more robust against the misclassification of causal and non-causal variants than the other collapsing methods. In addition, the CMC test has the advantage of allowing both rare and common variants to contribute to the overall test for the effect of a gene, although a large number of degrees of freedom are required when testing many common variants.

1.2.2 Weighted approach: WSS, KBAC

The CMC test depends on the ad hoc choice of a frequency cutoff to distinguish rare and common variants. The weighted approach, on the other hand, aggregates rare and common variants and assigns different weights to each group. In the weighted approach, the genetic variable for an individual is calculated as a weighted sum of the mutation counts. This approach accentuates signals from rare mutations such that the test is not completely dominated by common mutations. The Madsen and Browning method weights each mutation according to its frequency in the unaffected subjects and permutes the disease status to assess the significance of a Wilcoxon-type test statistic (Madsen and Browning 2009). In the Kernel-Based Adaptive Cluster (KBAC) method, the weight is based on the kernel functions, depending on the estimated sample risk, and the permutation procedure is applied to evaluate the significance of the score test statistics under a logistic regression model (Liu and Leal 2010).

1.2.3 Maximization approach: VT

The optimal choice of the MAF cutoff depends on the true disease model, which is unknown. In addition, a variant with frequency 0.01 is rare in a small data set of 500 individuals but is quite common in a much larger data set of 100,000 individuals. Therefore, a fixed-threshold may not be appropriate for all diseases and data sets. The variable threshold (VT) test developed by Price et al. (2010) uses the maximum of the test statistics over all unique allele-frequency thresholds and assesses statistical significance by permutation. This approach can be generalized to include multiple allele-frequency thresholds and different weight functions.

1.2.4 Signed approach: Han and Pan

The foregoing tests do not have good power if the variants being combined have opposite effects on the phenotype. Several other tests aim to detect variants with opposite effects. The methods proposed by Han and Pan incorporate the signs of the observed effects into the calculation of the genetic variables and apply a permutation procedure to assess the significance (Han and Pan 2010). This test was motivated by the data-adaptive modifications to an aggregation test originally proposed for common variant analysis, which aims to strike a balance between utilizing information from multiple markers in linkage disequilibrium and reducing the cost of large degrees of freedom or adjustments for multiple testing.

1.2.5 Variance-Component (VC) tests: Similarity regression, C-alpha, SKAT

VC tests are aimed at detecting variants with opposite effects within a gene. VC tests can be motivated from the similarity regression or kernel-machine regression. In these regression frameworks, the genetic effects are incorporated into the model through a nonparametric function $h(G_{i1}, \ldots, G_{iK})$, where G_{ik} is the genotype of the kth variant for the ith subject. Supposed $G_i = (G_{i1}, \ldots, G_{iK})$, then the form of the nonparametric function is determined by a user-specified, positive, semi-definite kernel matrix $K(G_i, G_j)$, which measures the genomic similarity between the genotypes of the ith and jth subjects. Some commonly used kernels include (weighted) linear, identity-by-descent, and quadratic kernels. By the representation theory, $h(G_i)$ can be written as $\sum_{j=1}^{n} \alpha_j K(G_i, G_j)$ with parameters $\alpha_1, \ldots, \alpha_n$. It can be shown that this nonparametric regression framework is equivalent to the random-effects model by treating h as subject-specific random effects with mean 0 and covariance matrix

τK. Therefore, testing $h = 0$ is equivalent to testing $\tau = 0$. Based on the working random-effects model, the score test statistic can be constructed, and the asymptotic distribution can be derived. A related test is SKAT-O, which is a weighted sum of the burden and VC statistics (Lee et al. 2012).

1.3 Trait-dependent sampling

The rare variants that are involved in complex trait etiologies usually only have moderate effect sizes, and even their aggregated frequencies across a genetic region can be limited. Therefore, a large number of samples must be sequenced and analyzed in order to have adequate power to detect associations. Although NGS is much more cost-effective than Sanger sequencing, it is still expensive to generate high read depth data for a large number of samples. To reduce expenses and increase statistical power of association tests, many sequencing projects select samples based on the value of a trait of primary interest (i.e., the primary trait). Indeed, previous research shows that trait-dependent sampling can substantially increase power comparing to random sampling with equal sample sizes. In the NHLBI ESP, multiple studies were included, each of which was focused on one primary trait. Subjects with extreme high or low values of the quantitative primary traits BMI, LDL and BP were selected for whole-exome sequencing. For the binary primary traits myocardial infarction (MI) and stroke, case-control (MI) and case-only (stroke) samples were generated for sequencing. In addition, a large random sample was created and referred to as the deeply phenotype reference (DPR). In addition to the primary trait, there were many quantitative and binary secondary traits across the six studies (e.g, high-density lipoprotein, triglyceride, diabetes). A mega- or meta-analysis that includes these secondary traits would boost statistical power.

However, the trait-dependent sampling study design produces challenges for analyzing secondary traits. If the secondary trait is correlated with the primary trait, and the primary trait is associated with a genetic variable, then spurious secondary trait association will be created among the subjects with extreme values of the primary trait. Thus, standard methods that ignore the sample ascertainment yield biased effect estimates and inflated type I error in association tests. For quantitative secondary traits, the properties of the naïve methods have been investigated theoretically and empirically (Lin et al. 2013). In the subsequent materials, we show the likelihood-based methods and the three types of gene-level tests that have been developed.

Suppose that we have a cohort of n subjects, among whom n_1 subjects are selected for sequencing. We assume that the primary trait Y_1 is available on all n cohort members. (If there are missing values on Y_1, we define n as the total number of subjects with available Y_1.) The selection of subjects for sequencing may depend on the values of Y_1 in the entire cohort. By definition, the genotype G is available only on the n_1 sequenced subjects. We assume that the covariate Z and the secondary trait Y_2 are available only on the n_1 sequenced subjects. The values of Y_2 may be missing among the sequenced subjects.

We allow G and Z to differ between the primary and secondary traits. The observed-data likelihood can be expressed as

$$\prod_{i=1}^{n_1} P(Y_{1i}|G_{1i}, Z_{1i})P(G_{1i}, Z_{1i}) \prod_{i=n_1+1}^{n} \sum_{g,z} P(Y_{1i}|g, z)P(g, z) \prod_{i=1}^{n_2} P(Y_{2i}|Y_{1i}, G_{2i}, Z_{2i}). \quad (1.1)$$

It is natural to formulate the joint distribution of Y_1 and Y_2 through the bivariate linear regression model:

$$Y_1 = \beta_1^{\mathrm{T}} G_1 + \gamma_1^{\mathrm{T}} Z_1 + \epsilon_1, \quad Y_2 = \beta_2^{\mathrm{T}} G_2 + \gamma_2^{\mathrm{T}} Z_2 + \epsilon_2,$$

where G_1 and G_2 may pertain to individual variants or (weighted) burden scores. Conditioned on Y_1, Y_2 satisfies the following linear regression model:

$$Y_2 = \delta \widetilde{Y}_1 + \beta_2^{\mathrm{T}} G_2 + \gamma_2^{\mathrm{T}} Z_2 + \widetilde{\epsilon}_2,$$

where $\widetilde{Y}_1 = Y_1 - \beta_1^{\mathrm{T}} G_1 - \gamma_1^{\mathrm{T}} Z_1$.

1.3.1 Estimating parameters for the primary trait

We maximize the first two terms in expression 1.1 to obtain the maximum likelihood estimates (MLEs) of $(\beta_1, \gamma_1, \sigma_{11})$ and $P(\cdot, \cdot)$. We adopt the nonparametric maximum likelihood estimation (NPMLE) approach to estimate $P(\cdot, \cdot)$ by the discrete probabilities at $(g_1, z_1), \cdots, (g_m, z_m)$, which are the distinct observed values of (G_{1i}, Z_{1i}) $(i = 1, \ldots, n_1)$. We denote the point mass at (g_j, z_j) as q_j. Then, we maximize the following objective function

$$\sum_{i=1}^{n_1} \left[\log P(Y_{1i}|G_{1i}, Z_{1i}) + \log \sum_{j=1}^{m} I\{(G_{1i}, Z_{1i}) = (g_j, z_j)\} q_j \right] + \sum_{i=n_1+1}^{n} \log \sum_{j=1}^{m} P(Y_{1i}|g_j, z_j) q_j,$$

where $I\{\cdot\}$ is the indicator function.

The maximization is carried out through an expectation-maximization (EM) algorithm, in which the missing values of (G_1, Z_1) for the non-sequenced subjects are inferred from the discrete probability distribution with point mass q_j at (g_j, z_j) $(j = 1, \cdots, m)$. Start with the initial values:

$$\beta_1 = 0, \quad \gamma_1 = 0, \quad \sigma_{11} = \text{sample variance of } Y_1 \text{ based on } (Y_{11}, \ldots, Y_{1n}),$$

$$\text{and} \quad q_j = 1/m, j = 1, \ldots, m.$$

We iterate between the following E-step and M-step until convergence.

E-step. For $i = 1, \ldots, n_1$, we set $\psi_{ij} = I\{(G_{1i}, Z_{1i}) = (g_j, z_j)\}$. For $i = n_1 + 1, \ldots, n$, we set

$$\psi_{ij} = \frac{P(Y_{1i}|g_j, z_j)q_j}{\sum_{k=1}^m P(Y_{1i}|g_k, z_k)q_k},$$

where $P(y_1|g, z) = (2\pi\sigma_{11})^{-1/2} \exp\{-(y_1 - \beta_1^{\mathrm{T}} g - \gamma_1^{\mathrm{T}} z)^2/(2\sigma_{11})\}$.

M-step. We update the parameter values as follows:

$$\eta = \left(\sum_{i=1}^n \sum_{j=1}^m \psi_{ij} W_j W_j^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^n Y_{1i} \sum_{j=1}^m \psi_{ij} W_j \right) \quad \text{and}$$

$$\sigma_{11} = n^{-1} \sum_{i=1}^n \sum_{j=1}^m \psi_{ij} (Y_{1i} - \eta^{\mathrm{T}} W_j)^2,$$

where

$$\eta = \begin{bmatrix} \beta_1 \\ \gamma_1 \end{bmatrix} \quad \text{and} \quad W_j = \begin{bmatrix} g_j \\ z_j \end{bmatrix}.$$

In addition,

$$q_j = n^{-1} \sum_{i=1}^n \psi_{ij}, \quad j = 1, \cdots, m.$$

At convergence, we obtain the estimator $(\widehat{\beta}_1, \widehat{\gamma}_1, \widehat{\sigma}_{11}, \widehat{q}_1, \ldots, \widehat{q}_m)$. It follows from Theorem 1 of Lin and Zeng (2006) that the estimator is consistent, asymptotically normal ,and asymptotically efficient. We estimate the asymptotic covariance matrix according to the Louis formula. For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, let l_{1ij} and l_{2ij} be the first and second derivatives, respectively, of $\log P(Y_{1i}|g_j, z_j) + \log q_j$ with respect to $(\beta_1, \gamma_1, \sigma_{11}, q_1, \ldots, q_m)$. We then calculate the information matrix as

$$Q_1 = -\sum_{i=1}^n \sum_{j=1}^m \psi_{ij} l_{2ij} - \sum_{i=1}^n \left\{ \sum_{j=1}^m \psi_{ij} l_{1ij} l_{1ij}^{\mathrm{T}} - \left(\sum_{j=1}^m \psi_{ij} l_{1ij} \right) \left(\sum_{j=1}^m \psi_{ij} l_{1ij} \right)^{\mathrm{T}} \right\}.$$

To account for the constraint that $\sum_{j=1}^m q_j = 1$, we let D denote the derivative matrix of $(\beta_1, \gamma_1, \sigma_{11}, q_1, \cdots, q_m)$ with respect to $(\beta_1, \gamma_1, \sigma_{11}, q_1, \cdots, q_{m-1})$. Then, the asymptotic covariance matrix of the estimator $(\widehat{\beta}_1, \widehat{\gamma}_1, \widehat{\sigma}_{11}, \widehat{q}_1, \cdots, \widehat{q}_{m-1})$ is estimated by $\Omega_1 = F^{-1}$, where $F = D^{\mathrm{T}} Q_1 D$.

1.3.2 Estimating parameters for the secondary trait

To estimate $(\delta, \beta_2, \gamma_2, \widetilde{\sigma}_{22})$, we maximize the last term in (2.1) or equivalently apply the standard least-squares method to the observations $(Y_{2i}, \widehat{Y}_{1i}, G_{2i}, Z_{2i})$ $(i = 1, \ldots, n_2)$, where $\widehat{Y}_{1i} = Y_{1i} - \widehat{\beta}_1^{\mathrm{T}} G_{1i} - \widehat{\gamma}_1^{\mathrm{T}} Z_{1i}$. That is,

$$\begin{bmatrix} \widehat{\delta} \\ \widehat{\beta}_2 \\ \widehat{\gamma}_2 \end{bmatrix} = \left(\sum_{i=1}^{n_2} \begin{bmatrix} \widehat{Y}_{1i} \\ G_{2i} \\ Z_{2i} \end{bmatrix}^{\otimes 2} \right)^{-1} \left(\sum_{i=1}^{n_2} Y_{2i} \begin{bmatrix} \widehat{Y}_{1i} \\ G_{2i} \\ Z_{2i} \end{bmatrix} \right)$$

and

$$\widehat{\widetilde{\sigma}}_{22} = n_2^{-1} \sum_{i=1}^{n_2} (Y_{2i} - \widehat{\delta}\widehat{Y}_{1i} - \widehat{\beta}_2^{\mathrm{T}} G_{2i} - \widehat{\gamma}_2^{\mathrm{T}} Z_{2i})^2,$$

where $a^{\otimes 2} = aa^{\mathrm{T}}$. We estimate the covariance matrix of $(\widehat{\delta}, \widehat{\beta}_2, \widehat{\gamma}_2)$ by

$$\Omega_2 = \widehat{\widetilde{\sigma}}_{22} \left(\sum_{i=1}^{n_2} \begin{bmatrix} \widehat{Y}_{1i} \\ G_{2i} \\ Z_{2i} \end{bmatrix}^{\otimes 2} \right)^{-1} + J \widetilde{\Omega}_1 J^{\mathrm{T}},$$

where J is the Jacobian matrix of $(\widehat{\delta}, \widehat{\beta}_2, \widehat{\gamma}_2)$ with respect to $(\widehat{\beta}_1, \widehat{\gamma}_1)$, and $\widetilde{\Omega}_1$ is the block of Ω_1 corresponding to (β_1, γ_1).

1.3.3 Performing association tests

To calculate the score statistic for testing the null hypothesis $H_0^{(1)} : \beta_1 = 0$, we calculate the restricted MLE of $(\gamma_1, \sigma_{11}, q_1, \cdots, q_m)$ under $H_0^{(1)}$. This is accomplished through the aforementioned EM algorithm, in which β_1 is set to 0 and only $(\gamma_1, \sigma_{11}, q_1, \cdots, q_m)$ is estimated. The score statistic for testing $H_0^{(1)} : \beta_1 = 0$ is

$$U_1 = \sum_{i=1}^{n} \sum_{j=1}^{m} \psi_{ij} l_{1ij}^{(1)},$$

where $l_{1ij}^{(1)}$ is the component of l_{1ij} corresponding to β_1. It can be shown that U_1 is approximately multivariate normal with mean 0 and covariance matrix

$$V_1 = F_{11} - F_{12} F_{22}^{-1} F_{21},$$

where $\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}$ is the partition of F for β_1 versus the other parameters.

We test $H_0^{(1)}$ using the quadratic form

$$U_1^{\mathrm{T}} V_1^{-1} U_1,$$

which is referred to the χ_d^2 distribution, where d is the dimension of β_1. We also consider the maximum test statistic

$$T_{\max} = \max_{j=1,\ldots,d} |T_j|,$$

where $T_j = U_{1j}/V_{1j}^{1/2}$, U_{1j} is the jth component of U_1, and V_{1j} is the jth diagnonal element of V_1. The p-value of $T_{\max}$ is determined by the multivariate normal distribution of U_1. Finally, we consider the weighted quadratic form

$$Q = U^{\mathrm{T}} W U,$$

where W is a diagonal matrix of weights that depend on the MAFs through a beta function. The null distribution of Q is approximated by $\sum_{j=1}^{d} \lambda_j \chi_{1,j}^2$, where $(\lambda_1, \ldots, \lambda_d)$ are the eigenvalues of $V_1^{1/2} W V_1^{1/2}$, and $(\chi^2_{1,1}, \ldots, \chi^2_{1,d})$ are independent χ^2_1 random variables.

To test the null hypothesis $H_0^{(2)} : \beta_2 = 0$, we estimate δ, γ_2 and $\widetilde{\sigma}_{22}$ under $\beta_2 = 0$:

$$\begin{bmatrix} \widehat{\delta} \\ \widehat{\gamma}_2 \end{bmatrix} = \left(\sum_{i=1}^{n_2} \begin{bmatrix} \widehat{Y}_{1i} \\ Z_{2i} \end{bmatrix}^{\otimes 2} \right)^{-1} \left(\sum_{i=1}^{n_2} Y_{2i} \begin{bmatrix} \widehat{Y}_{1i} \\ Z_{2i} \end{bmatrix} \right)$$

and

$$\widehat{\widetilde{\sigma}}_{22} = n_2^{-1} \sum_{i=1}^{n_2} (Y_{2i} - \widehat{\delta}\widehat{Y}_{1i} - \widehat{\gamma}_2^{\mathrm{T}} Z_{2i})^2.$$

We calculate

$$U_2 = \widehat{\widetilde{\sigma}}_{22}^{-1} \sum_{i=1}^{n_2} (Y_{2i} - \widehat{\delta}\widehat{Y}_{1i} - \widehat{\gamma}_2^{\mathrm{T}} Z_{2i}) G_{2i},$$

which is approximately multivariate normal with mean 0 and covariance matrix

$$V_2 = \widehat{\widetilde{\sigma}}_{22}^{-1} \left\{ \sum_{i=1}^{n_2} G_{2i}^{\otimes 2} - \left(\sum_{i=1}^{n_2} G_{2i} \begin{bmatrix} \widehat{Y}_{1i} \\ Z_{2i} \end{bmatrix}^{\mathrm{T}} \right) \left(\sum_{i=1}^{n_2} \begin{bmatrix} \widehat{Y}_{1i} \\ Z_{2i} \end{bmatrix}^{\otimes 2} \right)^{-1} \left(\sum_{i=1}^{n_2} \begin{bmatrix} \widehat{Y}_{1i} \\ Z_{2i} \end{bmatrix} G_{2i}^{\mathrm{T}} \right) \right\}$$

$$+ B \widetilde{\Omega}_1 B^{\mathrm{T}},$$

where B is the Jacobian matrix of U_2 with respect to $(\widehat{\beta}_1, \widehat{\gamma}_1)$. We test $H_0^{(2)}$ by using the aforementioned three types of test statistics.

1.4 Meta-analysis

1.4.1 Meta-analysis of GWAS

GWAS have successfully identified common SNPs with small to modest effects for virtually every complex human disease (Hardy and Singleton 2009). For variants with small effect sizes, the signal in a single study may be too weak to detect due to the small sample size. Meta-analysis is an important tool to combine evidence from multiple studies and explain part of the missing heritability that was not easy to capture in individual studies. Many new findings have been made through meta-analysis of GWAS (Saxena et al. 2007, Scott et al. 2009, Franke et al. 2010). Fixed-effects models have been adopted almost exclusively; these models assume a common genetic effect across studies. Multiple meta-analysis methods for GWAS have been developed and are described below.

1. *p-value-based methods*

Let p_k denote the p-value from the kth study among a total of K studies. Assuming that the K studies are independent, the simplest meta-analysis approach is to combine the p-values using Fisher's method

$$T = -2\sum_{k=1}^{K} log(p_k)$$

or Stouffer's method

$$Z = \frac{\sum_{k=1}^{K} Z_k w_k}{\sqrt{\sum_{k=1}^{K} w_k^2}},$$

where w_k is the square root of the sample size of the kth study, and

$$Z_k = (\text{sign for effect direction}) * \Phi^{-1}\left(1 - \frac{p_k}{2}\right),$$

where $\Phi(\cdot)$ is the cumulative density function for the standard normal distribution. The major disadvantage of the p-value-based meta-analysis method is that it can not provide an overall estimate of the effect size.

2. *Effect estimates*

Let $X_1, \cdots, X_K$ denote the estimate of the effect sizes for the K studies. Inverse-variance weighting is usually applied to estimate overall genetic effect under fixed-effects models. The overall effect size estimate takes the form

$$X = \frac{\sum_{k=1}^K w_k X_k}{\sum_{k=1}^K w_k},$$

where w_k is the inverse of the variance for X_k. The variance of the estimate is

$$\mathrm{var}(X) = \frac{1}{\sum_{k=1}^K w_k}.$$

3. *Bayesian methods*

Some consortia have applied Bayesian approaches for meta-analysis. For example, the Wellcome Trust Case Control Consortium has used the Bayes factor that represents the ratio of the probability of the data under the null hypothesis to the probability of the data under the alternative hypothesis. Bayesian models are intuitive, but the result may depend on the assumptions about the prior distribution, and the genome-wide implementation can be computationally intensive. It has been shown that Bayes factors and p-values often yield similar rankings for common variants. However, differences can be observed for rare or low-frequency variants (Wakefield 2009).

4. *Methods that account for between-study heterogeneity*

An alternative approach is the random-effects model in which the genetic effect is allowed to be heterogenous among studies. There are multiple sources of heterogeneity across genetic studies: populations with different demographic features (Waters et al. 2010, Heid et al. 2010); phenotypes with different definitions or measurements (Heid et al. 2009, Tobacco and Consortium 2010); and inconsistencies in data collection and manipulation (Ioannidis et al. 2007) (e.g., genotyping platforms, quality control criteria, or imputation methods). Nevertheless, the random-effects model is seldom used in genetic association studies for two reasons: (1) The random-effects model requires a large number of studies to make valid asymptotic inference, but the number of available genetic association studies is usually small. (2) The conventional test under the random-effects model gives less significant p-values than the corresponding test under the fixed-effects model. In a recent paper by Han and Eskin (2011), this phenomenon was described, and the cause was discovered to be the implicit assumption of heterogeneity under the null hypothesis that the variant is not associated with the trait of interest. A new test was proposed in their paper that relaxes this assumption and thus, is more powerful than the test under the fixed-effects model when heterogeneity exists. To account for the fact that the number of studies is small, the Monte-Carlo method is used to obtain the p-value.

1.4.2 Meta-analysis of sequencing studies

We reviewed various gene-level tests in Section 1.1. Three major types of gene-level tests are the burden, VT, and VC tests. None of these tests is universally most powerful. The burden and VT tests can outperform the VC test when a large proportion of the variants are causal and harboring unidirectional effects, whereas the VC tests tend to be more powerful when a small proportion of the variants are causal and harboring bidirectional effects. Therefore, it is necessary to develop different meta-analysis methods for each of these tests. Of course, one can always combine p-values through the Fisher and Stouffer methods; software that implements this approach has been developed (e.g., RAREMETAL). However, it has been shown that meta-analysis based on p-values for rare variants loses efficiency. In addition, it is impossible to address the heterogeneity issue using p-values. Tang and Lin (2013) and Liu et al. (2014) have developed meta-analysis methods that combine score statistics across studies for gene-level tests under fixed-effects models. We review the fixed-effects methods in the subsequent materials.

Suppose that we are interested in d genetic variables, which may be the genotypes of individual variants or the burden scores for a gene. For each of K studies, we calculate the (multivariate) score statistic for testing the null hypothesis that none of the d genetic variables has any effect on the trait of interest, and we also calculate the corresponding information matrix. We sum the score statistics and information matrices over the K studies to obtain the overall score statistic $\mathbf{U}$ and overall information matrix $\mathbf{V}$. Note that $\mathbf{U}$ is a $d \times 1$ vector and $\mathbf{V}$ is a $d \times d$ matrix. Under the null hypothesis, $\mathbf{U}$ is (asymptotically) multivariate zero-mean normal with covariance matrix $\mathbf{V}$. It can be shown that $\mathbf{U}$ is the score statistic for the common genetic effects in the joint likelihood based on the original data of the K studies, allowing nuisance parameters to be different among the studies (Lin and Zeng, 2010). Thus, association testing based on $\mathbf{U}$ and $\mathbf{V}$ is equivalent to the joint analysis of the original data.

Given $\mathbf{U}$ and $\mathbf{V}$, we perform three types of multivariate tests, which encompass all commonly used rare-variant tests.

1. *Quadratic statistic*:

$$Q = \mathbf{U}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{U}.$$

Under the null hypothesis, Q is distributed as χ_d^2. If $\mathbf{U}$ pertains to a specific burden score, then Q is a burden test. If $\mathbf{U}$ pertains to the genotype values of common SNPs and the burden score of rare variants, then Q is the CMC test (Li and Leal, 2008).

2. *Maximum statistic*:

$$T_{\max} = \max_{j=1,\ldots,d} {U_j}^2/V_j,$$

where U_j is the jth component of $\mathbf{U}$, and V_j is the jth diagonal element of $\mathbf{V}$. The p-value of $T_{\max}$ is determined by the multivariate normal distribution of $\mathbf{U}$ (Lin and Tang, 2011). If the genetic variables consist of the burden scores at different MAF thresholds, then $T_{\max}$ is the VT test. If the genetic variables pertain to different types of burden scores, then $T_{\max}$ can be used to adjust for multiple testing with those burden scores.

3. *Weighted quadratic statistic*:

$$Q_w = \mathbf{U}^{\mathrm{T}}\mathbf{W}\mathbf{U},$$

where $\mathbf{W}$ is a weight matrix. The null distribution of Q_w is determined by $\sum_{j=1}^{d} \lambda_j \chi^2_{1,j}$, where λ_j is the jth eigenvalue of $\mathbf{V}^{1/2}\mathbf{W}\mathbf{V}^{1/2}$, and $\chi^2_{1,1}, \ldots, \chi^2_{1,d}$ are independent χ^2_1 random variables. If the genetic variables are the genotypes of individual SNPs, then Q_w becomes the SKAT or C-alpha test. For the SKAT test, $\mathbf{W}$ is a diagonal matrix that depends on the MAFs through a beta function; for the C-alpha test, $\mathbf{W}$ is an identity matrix.

These fixed-effects meta-analysis methods lose power if the genetic effects are heterogeneous among studies. Recently, Lee et al. (2013) proposed two test statistics to allow for heterogeneous effects:

$$\text{Het-SKAT} = \sum_{j=1}^{d}\sum_{k=1}^{K} w_{kj}^2 S_{kj}^2 \quad \text{and}$$

$$\text{Het-SKAT-O} = \varrho\Bigg(\sum_{j=1}^{d}\sum_{k=1}^{K} w_{kj} S_{kj}\Bigg)^2 + (1-\varrho)\sum_{j=1}^{d}\sum_{k=1}^{K} w_{kj}^2 S_{kj}^2,$$

where S_{kj} is the score statistic for testing the jth variant in the kth study, w_{kj} is a weight for the jth variant, and ϱ is chosen to minimize the p-value. Het-SKAT is essentially a test of heterogeneity at the variant level. This test will not have good power if the average effect size is large or the heterogeneity exhibits at the burden score instead of the variant level. Het-SKAT-O is a weighted sum of fixed-effects burden test (under the additive mode of inheritance) and Het-SKAT and thus is a joint test of the mean effect at the burden score level and the heterogeneity at the variant level. Het-SKAT-O will lose power if both the mean effects and heterogeneity exhibit at the burden score level or if both the mean effects and heterogeneity exhibit at the individual variant level. The p-values of Het-SKAT and Het-SKAT-O are based on asymptotic distributions. Consequently, the type I error may not be well-controlled, the burden scores can only be calculated under the additive mode of inheritance, and the same set of weights has to be used for the two components of Het-SKAT-O.

In Chapter 4, we investigate the performance of the aforementioned meta-analysis methods and compare them with proposed random-effects methods.

CHAPTER2: A GENERAL FRAMEWORK FOR DETECTING DISEASE ASSOCIATIONS WITH RARE VARIANTS IN SEQUENCING STUDIES

2.1 Introduction

In this chapter, we provide a general framework for association testing with rare variants that reflects the spirits of the existing methods but is statistically more powerful and computationally more efficient. Our framework covers all major study designs (i.e., case-control, cross-sectional, cohort and family studies) and all common phenotypes (e.g., binary and quantitative traits, and potentially censored ages at onset of disease) and allows any covariates (e.g., environmental factors and ancestry variables). The ability to accommodate covariates is critically important because population stratification is expected to be a more severe issue with rare variants than with common variants but may be corrected by including suitable ancestry variables (e.g., percentage of African ancestry or principal components for ancestry) in the association analysis. We combine information across multiple variant sites within a gene by taking a weighted sum of the mutation counts for each study subject and relate the combined information and covariates to disease phenotypes through appropriate regression models. We derive theoretically optimal weights that would produce the most powerful tests among all valid tests and develop the corresponding testing procedures. We employ score-type statistics, which are numerically stable even in the case of extremely rare variants and computationally fast even in the presence of covariates. We provide asymptotic normal approximation for both fixed and variable threshold methods and develop permutation and other resampling tests that can accommodate covariates. We investigate theoretically and numerically when normal approximation is appropriate and when resampling is required. We modify the popular methods of Madsen and Browning and Price et al. to enhance statistical power, avoid permutation and accommodate covariates. We construct data-adaptive test statistics that are powerful even when the combined rare mutations have opposite effects on the phenotype. The advantages of the new methods over the existing ones are demonstrated both analytically and empirically.

2.2 Methods

Suppose that a total of n subjects are genotyped on a total of m SNPs in a gene and that there are d covariates. Here, the word "gene" refers to the group of variants that will be collectively analyzed and may pertain to a subset of SNPs within a gene or to a region/pathway involving multiple genes; "covariates" may include non-genetic variables, such as age and smoking status, as well as ancestry variables, such as percentage of African ancestry and principal components for ancestry. For $i = 1, \ldots, n$, let Y_i be the phenotype value of the ith subject; for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, let X_{ji} denote the number of the rare mutation the ith subject carries at the jth SNP; for $i = 1, \ldots, n$ and $j = 1, \ldots, d$, let Z_{ji} denote the value of the jth covariate on the ith subject. We can write

$$X_i = \begin{bmatrix} X_{1i} \\ \vdots \\ X_{mi} \end{bmatrix}, \quad Z_i = \begin{bmatrix} 1 \\ Z_{1i} \\ \vdots \\ Z_{di} \end{bmatrix}.$$

We first focus on binary phenotypes and then consider other common phenotypes.

2.2.1 Binary phenotypes

It is natural to relate Y_i to X_i and Z_i through the logistic regression model:

$$\Pr(Y_i = 1) = \frac{e^{\beta^{\mathrm{T}} X_i + \gamma^{\mathrm{T}} Z_i}}{1 + e^{\beta^{\mathrm{T}} X_i + \gamma^{\mathrm{T}} Z_i}},$$

where β and γ are $m \times 1$ and $(d+1) \times 1$ vectors of unknown regression parameters. Since the first component of Z_i is 1, the first component of γ corresponds to the intercept. Write $\beta = \tau\xi$, where τ is a scalar constant, and $\xi = \beta/\tau$. Then equation above becomes

$$\Pr(Y_i = 1) = \frac{e^{\tau S_i + \gamma^{\mathrm{T}} Z_i}}{1 + e^{\tau S_i + \gamma^{\mathrm{T}} Z_i}},$$

where $S_i = \xi^{\mathrm{T}} X_i$. Note that $\xi = (\xi_1, \ldots, \xi_m)^{\mathrm{T}}$ is a $m \times 1$ vector of weights and that S_i is a weighted linear combination of $X_{1i}, \ldots, X_{mi}$ with X_{ji} receiving the weight ξ_j. We will refer to ξ as the weight function.

The score statistic for testing the null hypothesis $H_0 : \tau = 0$ takes the form

$$U = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\widehat{\gamma}^{\mathrm{T}} Z_i}}{1 + e^{\widehat{\gamma}^{\mathrm{T}} Z_i}} \right) S_i,$$

where $\widehat{\gamma}$ is the restricted maximum likelihood estimator of γ, which solves the equation

$$\sum_{i=1}^{n} \left(Y_i - \frac{e^{\gamma^{\mathrm{T}} Z_i}}{1 + e^{\gamma^{\mathrm{T}} Z_i}} \right) Z_i = 0.$$

The variance of U is estimated by

$$V = \sum_{i=1}^{n} v_i S_i^2 - \left(\sum_{i=1}^{n} v_i S_i Z_i \right)^{\mathrm{T}} \left(\sum_{i=1}^{n} v_i Z_i Z_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{n} v_i S_i Z_i \right),$$

where

$$v_i = \frac{e^{\widehat{\gamma}^{\mathrm{T}} Z_i}}{(1 + e^{\widehat{\gamma}^{\mathrm{T}} Z_i})^2}.$$

Under H_0, the test statistic $T = U/V^{1/2}$ is asymptotically standard normal. In the absence of covariates,

$$U = \sum_{i=1}^{n} (Y_i - \overline{Y}) S_i,$$

and

$$V = \overline{Y}(1 - \overline{Y}) \left\{ \sum_{i=1}^{n} S_i^2 - n^{-1} \left(\sum_{i=1}^{n} S_i \right)^2 \right\},$$

where $\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$.

The true value of the weight function $\xi = (\xi_1, \ldots, \xi_m)^{\mathrm{T}}$ is unknown and must be determined biologically or empirically. If we set $\xi_j = 1$ $(j = 1, \ldots, m)$, then T is a burden test, which counts the total number of rare mutations each subject carries over the m SNPs. If we believe that common variants are not associated with the phenotype, then we set $\xi_j = 0$ if $p_j > c$, where p_j is the minor allele frequency (MAF) of the jth SNP, and c is a given threshold. If we set $\xi_j = \{p_j(1 - p_j)\}^{-1/2}$ $(j = 1, \ldots, m)$, then the weight function is in the same vein as that of Madsen and Browning.

If the choice of the weight function ξ is not proportional to β or ξ is estimated from the data, then U is no longer the score statistic. However, it is simple to verify that the test statistic T is asymptotically standard normal under H_0 regardless of how ξ is determined. The only condition is that, if ξ is estimated from the data, the estimate converges to a constant vector as the sample size n increases. This condition is satisfied by all sensible estimates, including those based on estimated allele frequencies. If the choice of ξ or the limit of the estimate of ξ is proportional to β, then the corresponding test statistic T is the most powerful among all valid tests.

The weight function ξ is similar to that of Price et al. The latter authors showed that, for case-control studies with known allele frequencies in the control population, the choice of $\xi_j = \{p_j(1-p_j)\}^{-1/2}$ $(j = 1, \ldots, m)$ corresponds to the implicit assumption that $\log(OR_j) \propto \{p_j(1-p_j)\}^{-1/2}$ $(j = 1, \ldots, m)$, where OR_j is the odds ratio in the 2×2 table for the jth SNP. Our theory is much more general in that it assumes unknown allele frequencies and accommodates covariates. Indeed, the proposed test statistic is optimal if ξ is proportional to the set of regression parameters (in the limit); this result holds for all phenotypes, including binary and continuous traits, as well as potentially censored ages at onset of disease.

Madsen and Browning suggested to set $\xi_j = \{\widehat{p}_j(1-\widehat{p}_j)\}^{-1/2}$ $(j = 1, \ldots, m)$, where $\widehat{p}_j$ is the estimate of the MAF of the jth SNP in the unaffected subjects. Because the weights depends on the phenotype values, the authors suggested a permutation-based test. Our testing framework allows such data-dependent weights since the frequency estimates converge to the true values as n increases. To improve the accuracy of asymptotic approximation, we suggest to estimate the frequencies from all study subjects rather than the unaffected subjects. Because the variants can be very rare, we recommend to add pseudo counts when estimating the frequencies, as was done by Madsen and Browning. The weight functions based on the frequency estimates in the pooled sample and the unaffected subjects will be denoted by "MB_p" and "MB_u", respectively; the constant weight function will be denoted by "C". The corresponding tests will be referred to as "MB_p-test", "MB_u-test" and "C-test".

Although MB_u is the weight function used by Madsen and Browning, our MB_u-test is fundamentally different from the Madsen and Browning (MB) test. The latter is based on the sum of the ranks of the S_i's with weight function MB_u over the affected subjects. Madsen and Browning proposed to assess the statistical significance of their rank-sum statistic by permutation. They also suggested an asymptotic normal approximation by standardizing the rank-sum statistic by its mean and standard derivation.

Because the mean and standard derivation are estimated by permutation, the asymptotic version of the MB test is many orders of magnitudes slower than our asymptotic tests. The rank-sum statistic is confined to case-control analysis without covariates.

Price et al. developed a VT method by taking the maximum of the test statistics (i.e., z-scores) over all allele-frequency thresholds and assessing statistical significance by permutation. We describe below a more general approach that allows not only multiple allele-frequency thresholds but also different types of weight function; it also accommodates covariates and does not require permutation.

We consider K choices of ξ, which may correspond to different thresholds or different types of weight function, or both. For the kth choice of ξ, the corresponding S_i is denoted by S_{ki}. Then the "score" statistic is

$$U_k = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\widehat{\gamma}^{\mathrm{T}} Z_i}}{1 + e^{\widehat{\gamma}^{\mathrm{T}} Z_i}} \right) S_{ki},$$

and the test statistic is $T_k = U_k / V_k^{1/2}$, where

$$V_k = \sum_{i=1}^{n} v_i S_{ki}^2 - \left(\sum_{i=1}^{n} v_i S_{ki} Z_i \right)^{\mathrm{T}} \left(\sum_{i=1}^{n} v_i Z_i Z_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{n} v_i S_{ki} Z_i \right).$$

We can show that, under H_0, the random vector $(U_1, \ldots, U_K)^{\mathrm{T}}$ is approximately K-variate normal with mean 0 and covariance matrix $\{V_{kl}; k, l = 1, \ldots, K\}$, where

$$V_{kl} = \sum_{i=1}^{n} v_i S_{ki} S_{li} - \left(\sum_{i=1}^{n} v_i S_{ki} Z_i \right)^{\mathrm{T}} \left(\sum_{i=1}^{n} v_i Z_i Z_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{n} v_i S_{li} Z_i \right).$$

For the two-sided test, we consider the maximum of the absolute test statistics

$$T_{\max} = \max_{k=1,\ldots,K} |T_k|.$$

Let $t_{\max}$ be the observed value of $T_{\max}$. The p-value is given by

$$\Pr(T_{\max} \geq t_{\max}) = 1 - \Pr(|T_1| < t_{\max}, \ldots, |T_K| < t_{\max}),$$

which is evaluated by treating $(T_1, \ldots, T_K)^{\mathrm{T}}$ as a K-variate normal random vector with mean 0 and covariance matrix $\{r_{kl}; k, l = 1, \ldots, K\}$, where $r_{kl} = V_{kl}/(V_{kk}V_{ll})^{1/2}$. (The one-sided p-value can be calculated in a similar manner.) We reject H_0 if the p-value is smaller than the nominal significance level α.

The tests based on positive weight functions, such as C, MB_u and MB_p, will have low power if the mutations being combined have opposite effects on the phenotype. The optimal choice of ξ_j is β_j, which is unknown. We can estimate β_j from the data. It would be tempting to set ξ_j to $\widehat{\beta}_j$, where $\widehat{\beta}_j$ is an appropriate estimate of β_j. There are two major problems with this strategy. First, the test statistic T will not be asymptotically normal. Second, the $\widehat{\beta}_j$'s are highly variable (since the individual variants are very rare) and can be quite different from the true values of the β_j's. As a compromise, we set $\xi_j = \widehat{\beta}_j + \delta$, where δ is a given constant. We refer to this weight function as "EREC", an abbreviation of *e*stimated *re*gression *c*oefficients. The corresponding test statistic T will be asymptotically standard normal as long as δ is non-zero. Indeed, the EREC test is asymptotically optimal in that ξ_j will converge to β_j if we let δ decrease to 0 as the sample size n increases to ∞. The asymptotic normality and optimality require large samples. For small samples, we recommend to use a relatively large value of δ so that the weights are not unduly driven by the highly variable $\widehat{\beta}_j$'s.

The Han and Pan (HP) statistic is a special case of our score statistic U (for binary traits without covariates) in which $\xi_j = -1$ if $\widehat{\beta}_j < 0$ and the corresponding p-value < 0.1 and $\xi_j = 1$ otherwise. The SKAT statistic of Wu et al. is a weighted sum of the squared score statistics for individual variants, and the C-alpha statistic of Neale et al. is an unweighted sum for binary traits without covariates. Unlike the EREC test, the

HP, C-alpha and SKAT tests are not asymptotically optimal.

Because the asymptotic approximation may not be accurate in small samples, especially when the weight function ξ involves the phenotype values Y_i's, we also provide permutation-type tests. In the absence of covariates, we simply permute the phenotype values Y_i's and calculate the test statistic T for each permutation. Note that it is necessary to re-calculate the S_i's after permuting the Y_i's if the weight function ξ depends on the Y_i's.

Our permutation differs from that of Price et al. in that we permute T whereas they permuted $\sum_{i=1}^{n} Y_i S_i$. The former is a pivotal statistic whereas the latter is not. (It is desirable to permute a pivotal statistic.) If the test is one-sided and the weight function does not depend on the phenotype values, then our permutation is equivalent to Price et al.'s; otherwise, the two are different. For VT methods, the numerators in the z-scores of Price et al. are the same as ours, but the denominators are not the same as or proportional to ours. Thus, the permutation p-values are generally different between the two methods. The permutation version of the MB test requires ranking the S_i's for each permutation and is thus substantially slower than our permutation tests.

In the presence of covariates, it is not appropriate to permute the Y_i's because Y_i is generally correlated with Z_i. Instead, we generate Y_i^* from the fitted null model:

$$\Pr(Y_i^* = 1) = \frac{e^{\widehat{\gamma}^{\mathrm{T}} Z_i}}{1 + e^{\widehat{\gamma}^{\mathrm{T}} Z_i}},$$

and replace the Y_i's with the Y_i^*'s to calculate the test statistic. This process is repeated and is called (parametric) bootstrap. Both permutation and bootstrap are resampling methods. In the absence of covariates, $\Pr(Y_i^* = 1)$ is the sample proportion of cases.

A large number of resamples (i.e., permutations or bootstrap samples) are required to obtain an accurate estimate of a small p-value. However, most p-values are relatively large and can be estimated accurately with a small number of resamples. Thus, we employ a multi-stage procedure which filters out large p-values with small numbers of resamples and uses large numbers of resamples only for the most extreme p-values.

2.2.2 Quantitative phenotypes

For quantitative traits, we consider the linear regression model:

$$Y_i = \tau S_i + \gamma^{\mathrm{T}} Z_i + \epsilon_i,$$

where ϵ_i is normal with mean 0 and variance σ^2. Then the score statistic and its variance are

$$U = \sum_{i=1}^{n} \left(Y_i - \widehat{\gamma}^{\mathrm{T}} Z_i \right) S_i,$$

and

$$V = \widehat{\sigma}^2 \left\{ \sum_{i=1}^{n} S_i^2 - \left(\sum_{i=1}^{n} S_i Z_i \right)^{\mathrm{T}} \left(\sum_{i=1}^{n} Z_i Z_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{n} S_i Z_i \right) \right\},$$

where

$$\widehat{\gamma} = \left(\sum_{i=1}^{n} Z_i Z_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} Y_i Z_i,$$

and

$$\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \widehat{\gamma}^{\mathrm{T}} Z_i)^2.$$

For multiple weight functions,

$$U_k = \sum_{i=1}^{n} \left(Y_i - \widehat{\gamma}^{\mathrm{T}} Z_i \right) S_{ki},$$

and

$$V_{kl} = \widehat{\sigma}^2 \left\{ \sum_{i=1}^{n} S_{ki} S_{li} - \left(\sum_{i=1}^{n} S_{ki} Z_i \right)^{\mathrm{T}} \left(\sum_{i=1}^{n} Z_i Z_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{n} S_{li} Z_i \right) \right\}.$$

To perform permutation tests without covariates, we simply permute the Y_i's. In the presence of covariates, we permute the residuals $R_i = Y_i - \widehat{\gamma}^{\mathrm{T}} Z_i$ $(i = 1, \ldots, n)$ to yield the R_i^*'s, and replace Y_i by $Y_i^* = \widehat{\gamma}^{\mathrm{T}} Z_i + R_i^*$ $(i = 1, \ldots, n)$ in calculating the test statistic.

2.2.3 Survival outcomes

For potentially censored age-at-onset traits, we specify that the hazard function for the age at onset conditional on S_i and Z_i satisfies the proportional hazards model

$$\lambda(t|S_i, Z_i) = \lambda_0(t) e^{\tau S_i + \gamma^{\mathrm{T}} Z_i},$$

where λ_0 is an arbitrary baseline hazard function, and Z_i is redefined to exclude the unit component. Let Y_i denote the duration of follow-up for the ith subject, and let Δ_i indicate, by the values 1 vs 0, whether T_i is the actual age at onset or the censoring time. Then the score statistic and its variance are

$$U = \sum_{i=1}^{n} \Delta_i \left(S_i - \frac{\sum_{j \in \mathcal{R}_i} e^{\widehat{\gamma}^T Z_j} S_j}{\sum_{j \in \mathcal{R}_i} e^{\widehat{\gamma}^T Z_j}} \right),$$

and

$$V = \sum_{i=1}^{n} \Delta_i \left[\frac{\sum_{j \in \mathcal{R}_i} e^{\widehat{\gamma}^T Z_j} S_j^2}{\sum_{j \in \mathcal{R}_i} e^{\widehat{\gamma}^T Z_j}} - \left(\frac{\sum_{j \in \mathcal{R}_i} e^{\widehat{\gamma}^T Z_j} S_j}{\sum_{j \in \mathcal{R}_i} e^{\widehat{\gamma}^T Z_j}} \right)^2 \right],$$

where $\mathcal{R}_i$ denotes the set of subjects whose durations of follow-up are no shorter than Y_i, and $\widehat{\gamma}$ is the solution to the partial likelihood score equation

$$\sum_{i=1}^{n} \Delta_i \left(Z_i - \frac{\sum_{j \in \mathcal{R}_i} e^{\gamma^T Z_j} Z_j}{\sum_{j \in \mathcal{R}_i} e^{\gamma^T Z_j}} \right) = 0.$$

We observed through simulation studies that the asymptotic score test was anticonservative for the low-frequency variant. To tackle this problem, we adopt the parametric bootstrap to assess the statistical significance. The detailed steps are described below.

Let $S_T(t \mid z)$ and $S_C(t \mid z)$ denote the survival function for the event time and the censoring time respectively. For $i = 1, \ldots, n$,

1. generate event time T_i^* from the estimated survival function $S_T(t \mid z_i)$ and independently generate censoring time C_i^* from the estimated survival function $S_C(t \mid z_i)$;

2. set $Y_i^* = \min(T_i^*, C_i^*)$, with $\Delta_i^* = 1$ if $Y_i^* = T_i^*$ and $\Delta_i^* = 0$ otherwise.

The Breslow estimator of the cumulative hazard function can be written as

$$\hat{\Lambda}_0(t) = \sum_{i:Y_i \leq t} \frac{\Delta_i}{\sum\limits_{j \in \mathcal{R}_i} e^{Z_j \gamma}}$$

Survival function of the event time for the individual with covariate value z takes the form

$$\hat{S}_T(t \mid z) = \left(e^{-\hat{\Lambda}_0(t)} \right)^{e^{z\hat{\gamma}}}.$$

If the censoring time is independent of the covariate, the survival function of the censoring time can be estimated using the Kaplan-Meier estimator as (supposed no ties in

Y_i)

$$\hat{S}_C(t) = \prod_{i:Y_i \leq t} \left(\frac{n-i}{n+1-i} \right)^{1-\Delta_i}$$

If the censoring time is dependent on the covariate, the survival function of the censoring time can be estimated similarly based on the proportional hazards model.

2.2.4 Family-based data

Suppose that the study contains n families with n_i members in the ith family. For $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, let y_{ij}, s_{ij} and z_{ij} denote the values of Y, S and Z for the jth member of the ith family.

We consider using generalized estimation equations (GEE) models to capture the dependence of the trait values. Suppose that the marginal density of y_{ij} belongs to the exponential family with the form

$$P(y_{ij}) = \exp\left\{ (y_{ij}\theta_{ij} - b(\theta_{ij}))\phi + c(y_{ij}, \phi) \right\},$$

where $\theta_{ij} = \tau s_{ij} + \gamma^{\mathrm{T}} z_{ij}$, and $b(.)$ and $c(.)$ are specific functions. The mean and the variance of y_{ij} are given by

$$\mathrm{E}(y_{ij}) = b^{'}(\theta_{ij}) = \mu_{ij}, \quad \text{and} \quad \mathrm{Var}(y_{ij}) = \phi^{-1} b^{''}(\theta_{ij}),$$

where $b^{'}$ and $b^{''}$ are the first and second derivatives of the function $b(.)$.

Then we assume a working covariance matrix for $Y_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$

$$\phi^{-1} V_i = \phi^{-1} B_i^{1/2} R_i(\alpha) B_i^{1/2},$$

where $B_i = \mathrm{diag}(b^{''}(\theta_{i1}), \ldots, b^{''}(\theta_{in_i}))$, and $R_i(\alpha)$ is a working correlation matrix for Y_i with parameter α, which is usually assumed to be independent, exchangeable or proportional to the kinship matrix. Then, the set of GEEs is

$$S(\tau,\gamma,\alpha)=\begin{bmatrix}S_\tau(\tau,\gamma,\alpha)\\ S_\gamma(\tau,\gamma,\alpha)\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^n(\partial\mu_i/\partial\tau)^{\mathrm{T}}V_i^{-1}\left(Y_i-\mu_i\right)\\ \sum_{i=1}^n(\partial\mu_i/\partial\gamma)^{\mathrm{T}}V_i^{-1}\left(Y_i-\mu_i\right)\end{bmatrix},$$

where $\mu_i=(\mu_{i1},\ldots,\mu_{in_i})^{\mathrm{T}}$. We let $S_i(\tau,\gamma,\alpha)$ denote the term in $S(\tau,\gamma,\alpha)$ that the ith family contribute.

Suppose that $U=S_\mu(0,\widetilde{\gamma},\widetilde{\alpha})$, where $\widetilde{\gamma}$ and $\widetilde{\alpha}$ are the restricted maximum likelihood estimators (MLEs) of γ and α under H_0. The sandwich variance estimator of U takes this form

$$\Sigma=\begin{bmatrix}1 & -V_{\tau\gamma}V_{\gamma\gamma}^{-1}\end{bmatrix}\left[\sum_{i=1}^n\left(S_i(0,\widetilde{\gamma},\widetilde{\alpha})-S(0,\widetilde{\gamma},\widetilde{\alpha})/n\right)^{\otimes 2}\right]\begin{bmatrix}1 & -V_{\tau\gamma}V_{\gamma\gamma}^{-1}\end{bmatrix}^{\mathrm{T}},$$

where $V_{\tau\gamma}$ and $V_{\gamma\gamma}$ are the first derivatives of $S_\tau(\tau,\gamma,\alpha)$ and $S_\gamma(\tau,\gamma,\alpha)$ with respective to γ and then evaluated at the restricted MLEs.

We can compute the nuisance parameters γ and α through an iterative procedure described below:

Step 1. Compute an initial estimate $\gamma^{(0)}$ based on an independent working correlation matrix ($R(\alpha)=I$).

Step 2. Compute the working correlation matrix $R_i(\alpha)$ based on the Pearson residuals $(y_{ij}-\mu_{ij})/\sqrt{b''(\theta_{ij})}$ and the current $\gamma^{(l)}$.

Step 3. Compute $V_i=B_i^{1/2}R_i(\alpha)B_i^{1/2}$.

Step 4. Update γ according to

$$\gamma^{(l+1)}=\gamma^{(l)}+\left(D_i^{\mathrm{T}}V^{-1}D_i\right)^{-1}\sum_{i=1}^n D_i^{\mathrm{T}}V_i^{-1}(Y_i-\mu_i)\bigg|_{\gamma^{(l)}},$$

where $D_i = \begin{bmatrix} \partial\mu_i/\partial\tau & \partial\mu_i/\partial\gamma \end{bmatrix}$.

We iterate Step 2-4 until convergence. As an alternative, we can estimate $R_i(\alpha)$ as twice of the kinship matrix Φ_i. The element (k, l) in matrix Φ_i is defined as the expected proportion of genes shared identical by descent (IBD) by the jth and lth members within the ith family.

2.3 Simulation studies

We conducted extensive simulation studies to investigate the performance of the new and existing methods. We simulated case-control data with an equal number of cases and controls from model (1) in which the first component of γ was set to -2. We considered mainly the following six combinations of MAFs: (1) $p_j = 0.001j$ $(j = 1, \ldots, 10)$ with a total frequency of 5.5%; (2) $p_j = 0.0005j$ $(j = 1, \ldots, 10)$ with a total frequency of 2.75%; (3) $p_j = 0.00025j$ $(j = 1, \ldots, 20)$ with a total frequency of 5.25%; (4) $p_j = 0.005$ $(j = 1, \ldots, 10)$ with a total frequency of 5%; (5) $p_j = 0.0025$ $(j = 1, \ldots, 10)$ with a total frequency of 2.5%; and (6) $p_j = 0.0025$ $(j = 1, \ldots, 20)$ with a total frequency of 5%. The genotype values were simulated under Hardy-Weinberg equilibrium and linkage equilibrium. We did not use sophisticated population genetics models because we wished to control the number of variants and their frequencies, which allowed us to see clearly how the new and existing methods perform under various scenarios. We evaluated both asymptotic and resampling methods. When the simulation studies involved asymptotic methods only, we used 10 millions replicates (i.e., simulated data sets) to evaluate type I error and 100,000 replicates to evaluate power at $\alpha = 10^{-2}$, 10^{-3} and 10^{-4}. When the simulation studies involved resampling methods, we used 1 million replicates to evaluate type I error and 10,000 replicates to evaluate power at $\alpha = 10^{-2}$ and 10^{-3}. The resampling p-values were obtained from a 3-stage procedure with a maximum of 1 million resamples. The null hypothesis corresponded to $H_0 : \beta_j = 0$ $(j = 1, \ldots, m)$. We considered primarily two types of alternative hypotheses, $H_1 : \beta_j = x$ $(j = 1, \ldots, m)$ and $H_1 : \beta_j = x/\{p_j(1 - p_j)\}^{1/2}$ $(j = 1, \ldots, m)$, where x was chosen such that the power (of the most powerful method) was reasonably high at $\alpha = 10^{-2}$. We report below results from six series of simulation studies, the first four without covariates and the last two with covariates.

Our first series of simulation studies was designed to evaluate the new asymptotic methods with different weight functions. We considered the aforementioned six combinations of MAFs and generated data under the null hypothesis $H_0 : \beta_j = 0$ $(j = 1, \ldots, m)$, as well as two alternative hypotheses $H_1 : \beta_j = x$ $(j = 1, \ldots, m)$ and $H_1 : \beta_j = x/\{p_j(1 - p_j)\}^{1/2}$ $(j = 1, \ldots, m)$. We considered three (positive) weight functions: C, MB_p, and MB_u. We also considered the maximum of the test statistics based on weight functions C and MB_p, which will be referred to as $T_{\max}$. The results for the first combination of MAFs are displayed in Table 2.1, while those of the remaining five combinations are provided in Tables 2.7-2.11. The C-test, MB_p-test and $T_{\max}$ are conservative; the conservativeness decreases as n, α or total allele frequency increases. As expected, the C-test is more powerful than the MB_p-test under the first alternative hypothesis and less powerful under the second alternative hypothesis; $T_{\max}$ is nearly as powerful as the C-test under the first alternative and nearly as powerful as the MB_p-test under the second alternative. The MB_u-test is unacceptably liberal; therefore, we will not consider this asymptotic test any further.

Our second series of studies was devoted to comparisons of asymptotic and permutation methods. In addition to the new methods, we evaluated the asymptotic and permutation versions of the MB test, as well as the permutation method of Price et al. with weight function MB_u. We simulated data in the same manner as the first series of studies. The results for the first combination of MAFs are displayed in Table 2.2. Due to the discreteness of the test statistic, the permutation version of the C-test is more conservative than its asymptotic counterpart and consequently less powerful. The permutation MB_p-test and MB_u-test do not appear to be conservative; the former appears to be slightly more powerful than the latter. The MB test was designed for the second alternative hypothesis, for which the new asymptotic test based on weight function MB_p is more powerful than the asymptotic version of the MB test while the new permutation tests based on weight functions MB_p and MB_u are more powerful than the permutation version of the MB test. For weight function MB_u, the permutation test of Price et al. is less powerful than our permutation test.

In the third series of studies, we compared fixed and variable threshold methods. We simulated 11 SNPs with MAFs $p_j = 0.001j$ $(j = 1, \ldots, 10)$ and $p_{11} = 0.03$. We considered the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{11} = 0$, as well as two alternative hypotheses $H_1 : \beta_1 = \beta_2 = \ldots = \beta_{10} = x$, $\beta_{11} = 0$, and $H_1 : \beta_1 = \beta_2 = \ldots = \beta_{11} = x$. For fixed threshold methods, we considered the thresholds of 0.01 and 0.05; the corresponding tests are referred to as the T1 and T5 tests. For VT methods, we excluded the thresholds for which the total numbers of rare mutations were fewer than 10. As shown in Table 2.3, all the tests appear to be conservative, especially when n and α are small. The permutation T1 and T5 tests are more conservative than their asymptotic counterparts. In theory, T1 and T5 are the most powerful under the first and second alternatives, respectively. Because the frequency estimates for rare variants are highly variable, T1 turns out to be the least powerful among all the tests under the first alternative. The VT tests have good power under both alternatives, and the asymptotic and permutation versions have similar power. The permutation version of our VT test is slightly more powerful than that of Price et al.

In the fourth set of studies, we compared the C-test, MB_p-test and EREC test, as well as the HP, C-alpha and SKAT tests. Note that the last four tests were designed to detect variants with opposite effects. The EREC, HP and C-alpha tests were based on permutation whereas the SKAT was based on the Davies method. For the EREC test, we set $\xi_j = \widehat{\beta}_j + 1$, where $\widehat{\beta}_j$ is the estimate of the log odds ratio β_j (after adding a pseudo-count of 1 to each of the four cells in the 2×2 table). For the SKAT test, we used the default weighted linear kernel function. We set $p_j = 0.001j$ $(j = 1, \ldots, 10)$ and considered the null hypothesis $H_0 : \beta_j = 0$ $(j = 1, \ldots, 10)$ and four alternative hypotheses $H_1 : \beta_j = x$ $(j = 1, \ldots, 10)$, $H_1 : \beta_j = x/\{p_j(1 - p_j)\}^{1/2}$ $(j = 1, \ldots, 10)$, $H_1 : \beta_j = x$ $(j = 1, \ldots, 8)$, $\beta_9 = -x$, $\beta_{10} = -2x$, and $H_1 : \beta_j = x$ $(j = 1, \ldots, 9)$, $\beta_{10} = -x/2$. As shown in Table 2.4, SKAT is highly conservative, especially when n and α are small. The EREC test is slightly less powerful than the C-test and MB_p-test when the SNP effects are all positive but is much more powerful than the latter when there are opposite effects. The EREC test is always more powerful than the HP, C-alpha and SKAT tests.

The above four sets of studies contained no covariates. We also conducted extensive studies with covariates. We generated data in the same manner as before except that we added a normally distributed covariate whose mean is equal to the total number of rare mutations and whose variance is equal to 1 and we set its regression parameter to 0.3. Some key results are presented in Tables 2.5 and 2.6. The T1, T5, MB_p and VT tests appear to be conservative, especially when n and α are small, and their asymptotic and bootstrap versions have similar power. The EREC test has similar power to the C-test and MB_p test when all SNP effects are positive and is much more powerful than the latter when there are opposite effects. The EREC test is substantially more powerful than the SKAT regardless of the alternative.

Table 2.1: Type I error[a]and power of asymptotic methods with different weight functions

		$H_0 : \beta_j = 0$				$H_1 : \beta_j = x$				$\beta_j = x/\{p_j(1-p_j)\}^{1/2}$			
n	α	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u
500	10^{-2}	0.99	0.97	0.97	3.47	0.84	0.81	0.83	0.91	0.83	0.85	0.84	0.93
	10^{-3}	0.89	0.85	0.84	4.53	0.57	0.53	0.56	0.72	0.55	0.58	0.57	0.75
	10^{-4}	0.73	0.70	0.67	4.76	0.31	0.27	0.30	0.46	0.29	0.31	0.30	0.50
1000	10^{-2}	1.00	0.99	0.98	3.13	0.87	0.84	0.86	0.92	0.93	0.95	0.94	0.98
	10^{-3}	0.94	0.92	0.91	4.35	0.63	0.59	0.62	0.74	0.75	0.79	0.78	0.89
	10^{-4}	0.89	0.83	0.81	5.45	0.38	0.33	0.36	0.50	0.52	0.56	0.54	0.71
2000	10^{-2}	0.99	0.99	0.99	2.55	0.95	0.94	0.95	0.97	0.97	0.98	0.98	0.99
	10^{-3}	0.97	0.96	0.95	3.41	0.82	0.78	0.81	0.86	0.87	0.90	0.89	0.95
	10^{-4}	0.89	0.91	0.89	4.25	0.61	0.54	0.59	0.68	0.69	0.74	0.73	0.84
4000	10^{-2}	0.99	0.99	0.99	2.04	0.98	0.98	0.98	0.99	0.98	0.99	0.99	0.99
	10^{-3}	0.98	0.98	0.98	2.56	0.92	0.89	0.91	0.93	0.90	0.93	0.92	0.96
	10^{-4}	0.97	0.96	0.94	3.07	0.77	0.72	0.76	0.80	0.75	0.80	0.79	0.86

[a] divided by α.

Table 2.2: Type I error[a]and power of asymptotic and permutation methods

			Asymptotic			Permutation				
	n	α	C	MB_p	MB	C	MB_p	MB_u	Price[b]	MB
$H_0 : \beta_j = 0$	500	10^{-2}	0.99	0.98	0.98	0.71	1.02	1.02	1.01	1.00
		10^{-3}	0.89	0.87	0.89	0.62	0.99	1.01	0.99	1.01
	1000	10^{-2}	1.00	1.00	1.00	0.79	1.01	1.03	1.01	1.01
		10^{-3}	0.96	0.96	0.93	0.72	1.01	1.02	1.01	1.02
$H_1 : \beta_j = x$	500	10^{-2}	0.84	0.81	0.82	0.81	0.81	0.81	0.79	0.82
		10^{-3}	0.57	0.54	0.54	0.54	0.55	0.54	0.49	0.56
	1000	10^{-2}	0.86	0.84	0.85	0.85	0.84	0.84	0.82	0.85
		10^{-3}	0.63	0.58	0.60	0.60	0.59	0.58	0.53	0.60
$H_1 : \beta_j = x/\{p_j(1-p_j)\}^{1/2}$	500	10^{-2}	0.83	0.85	0.82	0.80	0.85	0.84	0.81	0.82
		10^{-3}	0.56	0.59	0.54	0.52	0.59	0.57	0.51	0.55
	1000	10^{-2}	0.93	0.95	0.92	0.92	0.95	0.94	0.93	0.92
		10^{-3}	0.75	0.80	0.73	0.73	0.80	0.77	0.74	0.74

[a] divided by α

[b] with weight function MB_u

Table 2.3: Type I error[a]and power of fixed and variable threshold methods

			Asymptotic			Permutation			
	n	α	T1	T5	VT	T1	T5	VT	Price[b]
$H_0 : \beta_j = 0$	500	10^{-2}	0.95	0.97	0.86	0.65	0.74	0.89	0.88
		10^{-3}	0.80	0.85	0.62	0.57	0.63	0.82	0.81
	1000	10^{-2}	1.00	1.00	0.92	0.76	0.83	0.94	0.93
		10^{-3}	0.90	0.95	0.70	0.66	0.72	0.87	0.87
$H_1 : \beta_1 = \cdots = \beta_{10} = x, \beta_{11} = 0$	500	10^{-2}	0.49	0.68	0.74	0.43	0.65	0.75	0.75
		10^{-3}	0.20	0.37	0.45	0.17	0.34	0.48	0.47
	1000	10^{-2}	0.59	0.70	0.76	0.55	0.68	0.77	0.77
		10^{-3}	0.30	0.41	0.48	0.27	0.38	0.50	0.50
$H_1 : \beta_1 = \cdots = \beta_{11} = x$	500	10^{-2}	0.39	0.88	0.78	0.34	0.86	0.79	0.78
		10^{-3}	0.14	0.65	0.51	0.12	0.62	0.54	0.52
	1000	10^{-2}	0.45	0.88	0.76	0.41	0.87	0.77	0.76
		10^{-3}	0.18	0.65	0.48	0.16	0.62	0.50	0.49

[a] divided by α

[b] VT method of Price et al.

Table 2.4: Type I error[a]and power of asymptotic and permutation tests for detecting potentially opposite effects

			Asymptotic			Permutation				
	n	α	C	MB_p	SKAT	C	MB_p	EREC	HP	C-alpha
$H_0 : \beta_j = 0$	500	10^{-2}	0.95	0.95	0.53	0.68	1.00	1.01	0.89	0.91
		10^{-3}	0.83	0.77	0.26	0.60	0.94	0.97	0.91	0.87
	1000	10^{-2}	0.99	0.98	0.75	0.77	1.02	1.02	0.97	0.96
		10^{-3}	0.97	0.95	0.57	0.73	1.02	1.04	1.01	0.97
$H_1 : \beta_j = x$	500	10^{-2}	0.77	0.74	0.33	0.73	0.74	0.72	0.71	0.36
		10^{-3}	0.49	0.45	0.09	0.46	0.47	0.44	0.41	0.14
	1000	10^{-2}	0.81	0.77	0.41	0.78	0.77	0.78	0.73	0.42
		10^{-3}	0.56	0.50	0.16	0.53	0.51	0.51	0.42	0.17
$H_1 : \beta_j = x/\{p_j(1-p_j)\}^{1/2}$	500	10^{-2}	0.76	0.78	0.26	0.73	0.79	0.71	0.70	0.27
		10^{-3}	0.47	0.50	0.06	0.44	0.51	0.41	0.39	0.08
	1000	10^{-2}	0.66	0.70	0.22	0.63	0.70	0.65	0.57	0.21
		10^{-3}	0.37	0.41	0.06	0.35	0.42	0.35	0.26	0.06
$H_1 : \beta_1 = \cdots = \beta_8 = x,$	500	10^{-2}	0.29	0.23	0.58	0.25	0.23	0.76	0.63	0.61
$\beta_9 = -x,\ \beta_{10} = -2x$		10^{-3}	0.09	0.06	0.25	0.08	0.06	0.49	0.38	0.32
	1000	10^{-2}	0.31	0.27	0.81	0.28	0.27	0.88	0.86	0.81
		10^{-3}	0.10	0.08	0.54	0.09	0.09	0.66	0.65	0.56
$H_1 : \beta_1 = \cdots = \beta_9 = x,$	500	10^{-2}	0.77	0.74	0.50	0.74	0.75	0.82	0.76	0.54
$\beta_{10} = -x/2$		10^{-3}	0.49	0.45	0.21	0.46	0.47	0.57	0.47	0.26
	1000	10^{-2}	0.86	0.85	0.69	0.84	0.85	0.92	0.86	0.70
		10^{-3}	0.64	0.61	0.40	0.61	0.62	0.73	0.60	0.42

[a] type I error is divided by the nominal significance level

Table 2.5: Type I error[a] and power of fixed and variable threshold methods with covariates

			Asymptotic				Bootstrap			
	n	α	T1	T5	MB_p	VT	T1	T5	MB_p	VT
$H_0 : \beta_j = 0$	500	10^{-2}	0.90	0.96	0.93	0.79	0.92	0.97	0.94	0.87
		10^{-3}	0.69	0.92	0.84	0.54	0.77	0.93	0.88	0.69
	1000	10^{-2}	0.95	0.98	0.97	0.80	0.97	0.95	0.98	0.87
		10^{-3}	0.80	0.93	0.85	0.60	0.86	0.81	0.92	0.76
$H_1 : \beta_1 = \ldots = \beta_{10} = x, \beta_{11} = 0$	500	10^{-2}	0.32	0.56	0.67	0.64	0.32	0.56	0.67	0.64
		10^{-3}	0.09	0.25	0.34	0.32	0.10	0.25	0.35	0.34
	1000	10^{-2}	0.41	0.60	0.71	0.68	0.41	0.60	0.71	0.69
		10^{-3}	0.16	0.30	0.41	0.37	0.16	0.30	0.41	0.39
$H_1 : \beta_1 = \ldots = \beta_{11} = x$	500	10^{-2}	0.22	0.76	0.71	0.65	0.22	0.75	0.71	0.65
		10^{-3}	0.05	0.47	0.40	0.34	0.05	0.47	0.40	0.36
	1000	10^{-2}	0.26	0.80	0.72	0.64	0.26	0.80	0.72	0.65
		10^{-3}	0.07	0.52	0.41	0.34	0.08	0.51	0.41	0.36

[a] divided by α

Table 2.6: Type I error[a]and power of asymptotic and bootstrap tests for detecting potentially opposite effects in the presence of covariates

			Asymptotic			Bootstrap		
	n	α	C	MB_p	SKAT	C	MB_p	EREC
$H_0 : \beta_j = 0$	500	10^{-2}	0.97	0.97	0.63	1.00	1.00	0.97
		10^{-3}	0.85	0.80	0.37	0.94	0.92	0.93
	1000	10^{-2}	0.98	0.97	0.81	0.99	0.99	0.98
		10^{-3}	1.01	0.96	0.56	1.05	1.01	0.99
$H_1 : \beta_j = x$	500	10^{-2}	0.67	0.63	0.14	0.67	0.63	0.67
		10^{-3}	0.37	0.33	0.02	0.37	0.33	0.37
	1000	10^{-2}	0.74	0.69	0.23	0.74	0.70	0.75
		10^{-3}	0.45	0.40	0.06	0.46	0.41	0.47
$H_1 : \beta_j = x/\{p_j(1-p_j)\}^{1/2}$	500	10^{-2}	0.65	0.68	0.32	0.65	0.68	0.65
		10^{-3}	0.35	0.37	0.08	0.36	0.38	0.35
	1000	10^{-2}	0.58	0.63	0.47	0.59	0.63	0.62
		10^{-3}	0.30	0.33	0.18	0.30	0.33	0.32
$H_1 : \beta_1 = \cdots = \beta_8 = x$,	500	10^{-2}	0.20	0.14	0.55	0.20	0.14	0.73
$\beta_9 = -x$, $\beta_{10} = -2x$		10^{-3}	0.05	0.03	0.23	0.06	0.03	0.44
	1000	10^{-2}	0.22	0.18	0.81	0.22	0.18	0.84
		10^{-3}	0.06	0.04	0.55	0.07	0.04	0.61
$H_1 : \beta_1 = \cdots = \beta_9 = x$,	500	10^{-2}	0.67	0.63	0.31	0.67	0.63	0.78
$\beta_{10} = -x/2$		10^{-3}	0.36	0.32	0.09	0.37	0.33	0.50
	1000	10^{-2}	0.79	0.76	0.53	0.79	0.77	0.89
		10^{-3}	0.51	0.48	0.23	0.52	0.49	0.67

[a] type I error is divided by the nominal significance level

Table 2.7: Type I error[a]and power of asymptotic methods with different weight functions under $p_j = 0.0005j$ $(j = 1, \cdots, 10)$

		$H_0 : \beta_j = 0$				$H_1 : \beta_j = x$				$\beta_j = x/\{p_j(1-p_j)\}^{1/2}$			
n	α	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u
500	10^{-2}	0.87	0.89	0.87	1.81	0.77	0.75	0.77	0.88	0.79	0.80	0.79	0.91
	10^{-3}	0.68	0.66	0.62	1.62	0.49	0.45	0.48	0.66	0.51	0.52	0.51	0.70
	10^{-4}	0.47	0.36	0.36	1.10	0.24	0.21	0.22	0.38	0.25	0.26	0.25	0.42
1000	10^{-2}	0.96	0.94	0.93	2.12	0.78	0.75	0.77	0.88	0.92	0.93	0.93	0.97
	10^{-3}	0.88	0.81	0.80	2.55	0.51	0.46	0.49	0.66	0.73	0.76	0.75	0.88
	10^{-4}	0.72	0.64	0.64	2.60	0.27	0.23	0.25	0.41	0.49	0.52	0.51	0.70
2000	10^{-2}	0.97	0.97	0.96	1.96	0.82	0.78	0.81	0.88	0.97	0.98	0.97	0.99
	10^{-3}	0.93	0.89	0.89	2.53	0.56	0.51	0.55	0.68	0.86	0.89	0.88	0.95
	10^{-4}	0.80	0.76	0.75	3.00	0.32	0.27	0.30	0.44	0.68	0.72	0.71	0.84
4000	10^{-2}	0.99	0.98	0.98	1.64	0.72	0.68	0.71	0.79	0.98	0.98	0.98	0.99
	10^{-3}	0.95	0.93	0.93	2.05	0.45	0.40	0.43	0.54	0.89	0.92	0.91	0.96
	10^{-4}	0.90	0.83	0.86	2.49	0.23	0.19	0.22	0.30	0.73	0.78	0.77	0.87

[a] divided by α

Table 2.8: Type I error[a]and power of asymptotic methods with different weight functions under $p_j = 0.00025j$ $(j = 1 \cdots, 20)$

		$H_0 : \beta_j = 0$				$H_1 : \beta_j = x$				$H_1 : \beta_j = x/\{p_j(1-p_j)\}^{1/2}$			
n	α	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u
500	10^{-2}	0.95	0.94	0.93	3.18	0.75	0.72	0.74	0.91	0.90	0.91	0.91	0.98
	10^{-3}	0.85	0.82	0.81	4.18	0.46	0.43	0.45	0.72	0.70	0.72	0.72	0.90
	10^{-4}	0.66	0.62	0.60	4.53	0.22	0.20	0.21	0.46	0.45	0.47	0.46	0.74
1000	10^{-2}	0.97	0.97	0.96	3.52	0.79	0.75	0.78	0.92	0.72	0.75	0.74	0.91
	10^{-3}	0.92	0.88	0.88	5.54	0.52	0.48	0.51	0.75	0.44	0.47	0.46	0.74
	10^{-4}	0.78	0.75	0.74	7.85	0.28	0.25	0.27	0.52	0.22	0.24	0.23	0.51
2000	10^{-2}	0.99	0.99	0.99	3.00	0.91	0.88	0.91	0.96	0.63	0.67	0.66	0.85
	10^{-3}	0.96	0.93	0.94	4.79	0.73	0.68	0.72	0.86	0.35	0.38	0.37	0.63
	10^{-4}	0.88	0.86	0.87	7.07	0.50	0.44	0.49	0.69	0.16	0.18	0.17	0.39
4000	10^{-2}	1.00	1.00	0.99	2.30	0.96	0.95	0.96	0.98	0.94	0.96	0.95	0.99
	10^{-3}	0.99	0.97	0.97	3.36	0.86	0.82	0.85	0.92	0.79	0.84	0.83	0.93
	10^{-4}	0.98	0.92	0.95	4.70	0.68	0.62	0.67	0.79	0.58	0.65	0.63	0.81

[a] divided by α

Table 2.9: Type I error[a]and power of asymptotic methods with different weight functions under $p_j = 0.005$ $(j = 1, \cdots, 10)$

		$H_0 : \beta_j = 0$				$H_1 : \beta_j = x$				$H_1 : \beta_j = x/\{p_j(1 - p_j)\}^{1/2}$			
n	α	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u
500	10^{-2}	0.94	0.94	0.93	2.34	0.72	0.70	0.71	0.85	0.79	0.78	0.79	0.90
	10^{-3}	0.82	0.81	0.80	2.83	0.43	0.41	0.42	0.62	0.51	0.50	0.51	0.70
	10^{-4}	0.62	0.60	0.58	2.86	0.20	0.19	0.19	0.36	0.27	0.25	0.26	0.44
1000	10^{-2}	0.98	0.98	0.97	2.13	0.76	0.75	0.76	0.86	0.91	0.91	0.91	0.96
	10^{-3}	0.92	0.91	0.91	2.94	0.49	0.48	0.49	0.64	0.74	0.72	0.73	0.84
	10^{-4}	0.81	0.74	0.76	3.76	0.26	0.25	0.25	0.40	0.50	0.48	0.49	0.65
2000	10^{-2}	0.98	0.98	0.98	1.61	0.89	0.89	0.89	0.94	0.97	0.96	0.97	0.98
	10^{-3}	0.96	0.96	0.96	2.06	0.70	0.70	0.70	0.79	0.86	0.86	0.86	0.91
	10^{-4}	0.90	0.91	0.91	2.56	0.47	0.46	0.46	0.58	0.68	0.67	0.68	0.78
4000	10^{-2}	0.99	0.99	0.99	1.29	0.96	0.96	0.96	0.97	0.98	0.98	0.98	0.99
	10^{-3}	0.97	0.97	0.97	1.48	0.84	0.83	0.84	0.88	0.90	0.90	0.90	0.93
	10^{-4}	0.94	0.92	0.92	1.67	0.65	0.64	0.64	0.72	0.74	0.74	0.74	0.81

[a] divided by α

Table 2.10: Type I error[a]and power of asymptotic methods with different weight functions under $p_j = 0.0025$ $(j = 1, \cdots, 10)$

		$H_0 : \beta_j = 0$				$H_1 : \beta_j = x$				$H_1 : \beta_j = x/\{p_j(1 - p_j)\}^{1/2}$			
n	α	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u
500	10^{-2}	0.85	0.88	0.86	1.85	0.82	0.81	0.82	0.92	0.82	0.81	0.82	0.92
	10^{-3}	0.67	0.64	0.62	1.58	0.56	0.54	0.55	0.73	0.56	0.54	0.55	0.74
	10^{-4}	0.41	0.36	0.32	0.92	0.29	0.27	0.28	0.46	0.29	0.27	0.28	0.46
1000	10^{-2}	0.97	0.94	0.94	2.36	0.93	0.93	0.93	0.97	0.94	0.93	0.93	0.97
	10^{-3}	0.86	0.82	0.81	2.91	0.78	0.76	0.77	0.89	0.78	0.76	0.77	0.89
	10^{-4}	0.66	0.64	0.62	3.00	0.55	0.53	0.54	0.72	0.55	0.53	0.54	0.72
2000	10^{-2}	0.97	0.97	0.96	2.13	0.91	0.91	0.91	0.96	0.98	0.97	0.97	0.99
	10^{-3}	0.91	0.89	0.89	2.94	0.73	0.72	0.72	0.84	0.89	0.89	0.89	0.94
	10^{-4}	0.84	0.82	0.82	3.64	0.49	0.48	0.49	0.65	0.73	0.72	0.72	0.84
4000	10^{-2}	0.99	0.99	0.99	1.61	0.90	0.90	0.90	0.94	0.98	0.98	0.98	0.99
	10^{-3}	0.95	0.95	0.94	2.06	0.71	0.70	0.70	0.80	0.92	0.91	0.92	0.95
	10^{-4}	0.91	0.93	0.91	2.62	0.47	0.47	0.47	0.59	0.78	0.78	0.78	0.86

[a] divided by α

Table 2.11: Type I error^{a}and power of asymptotic methods with different weight functions under $p_j = 0.0025$ $(j = 1, \cdots, 20)$

		$H_0 : \beta_j = 0$				$H_1 : \beta_j = x$				$H_1 : \beta_j = x/\{p_j(1-p_j)\}^{1/2}$			
n	α	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u	C	MB_p	$T_{\max}$	MB_u
500	10^{-2}	0.94	0.94	0.93	3.52	0.72	0.70	0.71	0.91	0.93	0.92	0.92	0.99
	10^{-3}	0.82	0.81	0.80	4.68	0.43	0.41	0.42	0.72	0.76	0.73	0.75	0.92
	10^{-4}	0.65	0.65	0.62	5.11	0.20	0.19	0.20	0.46	0.51	0.49	0.50	0.78
1000	10^{-2}	0.97	0.97	0.97	4.13	0.76	0.75	0.76	0.92	0.76	0.75	0.76	0.92
	10^{-3}	0.93	0.92	0.90	6.92	0.49	0.47	0.48	0.76	0.49	0.47	0.48	0.76
	10^{-4}	0.79	0.80	0.77	10.09	0.25	0.24	0.25	0.53	0.26	0.24	0.25	0.53
2000	10^{-2}	0.99	0.99	0.99	3.30	0.90	0.89	0.89	0.97	0.68	0.67	0.67	0.85
	10^{-3}	0.96	0.95	0.95	5.60	0.70	0.69	0.70	0.87	0.39	0.38	0.39	0.63
	10^{-4}	0.86	0.88	0.85	8.94	0.46	0.45	0.46	0.70	0.19	0.18	0.18	0.40
4000	10^{-2}	0.99	0.99	0.99	2.17	0.96	0.95	0.96	0.98	0.96	0.95	0.95	0.98
	10^{-3}	0.98	0.98	0.98	3.17	0.84	0.83	0.83	0.92	0.84	0.83	0.83	0.92
	10^{-4}	0.98	1.00	1.00	4.54	0.64	0.64	0.64	0.79	0.65	0.64	0.64	0.79

^{a} divided by α

Table 2.12: Type I error^{a}of score, Wald and LR tests with covariates

		$p_j = 0.001j$ $(j = 1, \cdots, 10)$						$p_j = 0.0005j$ $(j = 1, \cdots, 10)$					
		Score		Wald		LR		Score		Wald		LR	
n	α	C	MB_p	C	MB_p	C	MB_p	C	MB_p	C	MB_p	C	MB_p
500	10^{-2}	0.98	0.97	0.84	0.79	1.05	1.06	0.94	0.92	0.65	0.59	1.09	1.10
	10^{-3}	0.90	0.85	0.59	0.49	1.08	1.09	0.81	0.76	0.26	0.20	1.15	1.17
	10^{-4}	0.78	0.70	0.31	0.23	1.08	1.09	0.67	0.58	0.08	0.04	1.27	1.27
1000	10^{-2}	0.99	0.98	0.92	0.88	1.03	1.03	0.97	0.96	0.84	0.79	1.04	1.05
	10^{-3}	0.95	0.92	0.78	0.69	1.03	1.04	0.90	0.86	0.59	0.49	1.07	1.07
	10^{-4}	0.90	0.84	0.61	0.50	1.04	1.06	0.80	0.78	0.36	0.26	1.11	1.13

^{a} divided by α

Table 2.13: Power of score, Wald and LR tests with covariates under $H_1 : \beta_j = x$ $(j = 1, \ldots, 10)$

		$p_j = 0.001j\ (j = 1, \cdots, 10)$						$p_j = 0.0005j\ (j = 1, \cdots, 10)$					
		Score		Wald		LR		Score		Wald		LR	
n	α	C	MB_p	C	MB_p	C	MB_p	C	MB_p	C	MB_p	C	MB_p
500	10^{-2}	0.68	0.63	0.66	0.61	0.69	0.66	0.60	0.56	0.56	0.52	0.63	0.60
	10^{-3}	0.38	0.33	0.34	0.29	0.41	0.38	0.28	0.25	0.21	0.17	0.34	0.32
	10^{-4}	0.16	0.13	0.13	0.09	0.20	0.18	0.10	0.08	0.04	0.02	0.15	0.14
1000	10^{-2}	0.74	0.69	0.73	0.68	0.75	0.71	0.73	0.69	0.72	0.67	0.75	0.72
	10^{-3}	0.45	0.40	0.44	0.38	0.48	0.43	0.44	0.39	0.40	0.35	0.48	0.44
	10^{-4}	0.22	0.18	0.20	0.16	0.25	0.22	0.20	0.17	0.16	0.12	0.25	0.22

Table 2.14: Power of score, Wald and LR tests with covariates under $H_1 : \beta_j = x/\{p_j(1-p_j)\}^{1/2}$ $(j = 1, \ldots, 10)$

		$p_j = 0.001j\ (j = 1, \cdots, 10)$						$p_j = 0.0005j\ (j = 1, \cdots, 10)$					
		Score		Wald		LR		Score		Wald		LR	
n	α	C	MB_p	C	MB_p	C	MB_p	C	MB_p	C	MB_p	C	MB_p
500	10^{-2}	0.66	0.68	0.64	0.66	0.67	0.71	0.72	0.73	0.69	0.69	0.74	0.76
	10^{-3}	0.35	0.38	0.32	0.33	0.39	0.42	0.40	0.41	0.32	0.31	0.46	0.48
	10^{-4}	0.15	0.16	0.11	0.11	0.19	0.21	0.17	0.17	0.08	0.06	0.24	0.25
1000	10^{-2}	0.83	0.86	0.82	0.85	0.84	0.87	0.88	0.90	0.87	0.89	0.89	0.91
	10^{-3}	0.58	0.62	0.56	0.60	0.60	0.65	0.65	0.68	0.62	0.64	0.68	0.72
	10^{-4}	0.33	0.36	0.30	0.33	0.36	0.41	0.39	0.41	0.33	0.34	0.44	0.48

2.4 Data analysis

We considered high-depth sequence data from the exons of 202 genes encoding known or potential drug targets for 1,957 subjects randomly drawn from the CoLaus population-based collection. We analyzed total cholesterol (available in 1,899 subjects) as a quantitative trait and included eight covariates in the analysis: gender, age, age^2, and the top five principal components for ancestry constructed from the GWAS SNP data. One subject without the gender and age information was removed.

We restricted our analysis to polymorphic variants that are nonsense, missense or splice site mutations. We removed variants with observed MAFs>5% or missingness>10%. We excluded any gene whose total number of rare mutations is less than 5 and ended up with a total of 172 genes. There were a total of 2,304 variants in these 172 genes, and the number of variants per gene varied from 1 to 70, with a median of 11. We applied both the asymptotic and permutation versions of our T1, T5, MB_p and VT tests, as well as the permutation EREC test. We calculated the two-sided p-values. With 172 genes, the Bonferroni threshold at the 0.05 significance level corresponds to p-value of 0.0003 or $-\log_{10}(p\text{-value})$ of 3.5.

The results based on the asymptotic and permutation methods are shown in Figures 2.1 and 2.2, respectively. One gene was identified as the most significant by all the tests: the asymptotic p-values for T1, T5, MB_p and VT are 0.00011, 0.00011, 0.00021 and 0.00057, respectively; the corresponding permutation p-values are 0.00013, 0.00013, 0.00025, and 0.0012, respectively; the p-value of the EREC test is 0.00012. (The name of the gene is not disclosed here because the main study has not been published yet.) All the p-values, except the VT's, pass the Bonferroni criterion. Similar evidence of association has been observed in other samples of the sequencing project. There were 13 variants in the top gene. Their observed MAFs ranged from 0.00026 to 0.0024, the total frequency being 1.13%. Since the observed MAFs are all less than 1% in this case, T1 and T5 are the same test. For the VT test, the maximum occurs at the highest MAF. It is interesting to point out that common SNPs in the top gene were previously identified to be associated with total cholesterol.

We also performed a binary trait analysis by comparing high (i.e., > 6.2 mmol/L) and desirable (i.e., < 5.2 mmol/L) total cholesterol values. There were 451 subjects with high total cholesterol and 683 subjects with desirable total cholesterol. The results of the analysis are shown in Figures 2.3 and 2.4. All the tests identified the same top gene as in the case of the quantitative trait analysis: the asymptotic p-values for T1, T5, MB_p and VT are 0.00022, 0.00022, 0.00057 and 0.00088, respectively; the corresponding bootstrap p-values are 0.00019, 0.00019, 0.00039, and 0.00033, respectively. Again, T1 and T5 are the same test. The maximum of the VT test occurs at the highest MAF, at which threshold 18 out of the 451 subjects with high cholesterol values carry the rare mutations as opposed to 7 out of 683 subjects with desirable cholesterol values. The p-value of the bootstrap EREC test is 0.000021, which is the most extreme among all the tests and is even more extreme than all the p-values of the quantitative trait analysis. For eight out of the 10 variants in the top gene, there were more mutations in the high group than in the desirable group (17 vs 2); for the remaining two variants, there were fewer mutations in the high group than in the desirable group (1 vs 5). Thus, allowing opposite effects yielded stronger evidence of association than assuming effects of the same direction.

Finally, we compared the new methods with the existing ones. The results for the SKAT are shown in Figure 2.5 (top panel). For the top gene, the SKAT yielded the p-values of 0.0014 and 0.00024 in the quantitative and binary trait analyses, respectively, which are 10 times the p-values of our EREC test. Because the other existing methods do not allow covariates and some of them require binary traits, we performed the binary trait analysis without the covariates for all the methods. The results are shown in the bottom panel of Figure 2.5 and in Figures 2.6-2.8. Although the top gene remains the same, the results without covariate adjustment (for the top gene) are considerably less significant than those with covariate adjustment. For the top gene, the EREC test yielded a much more significant result (p-value= 0.00013) than all the other tests.

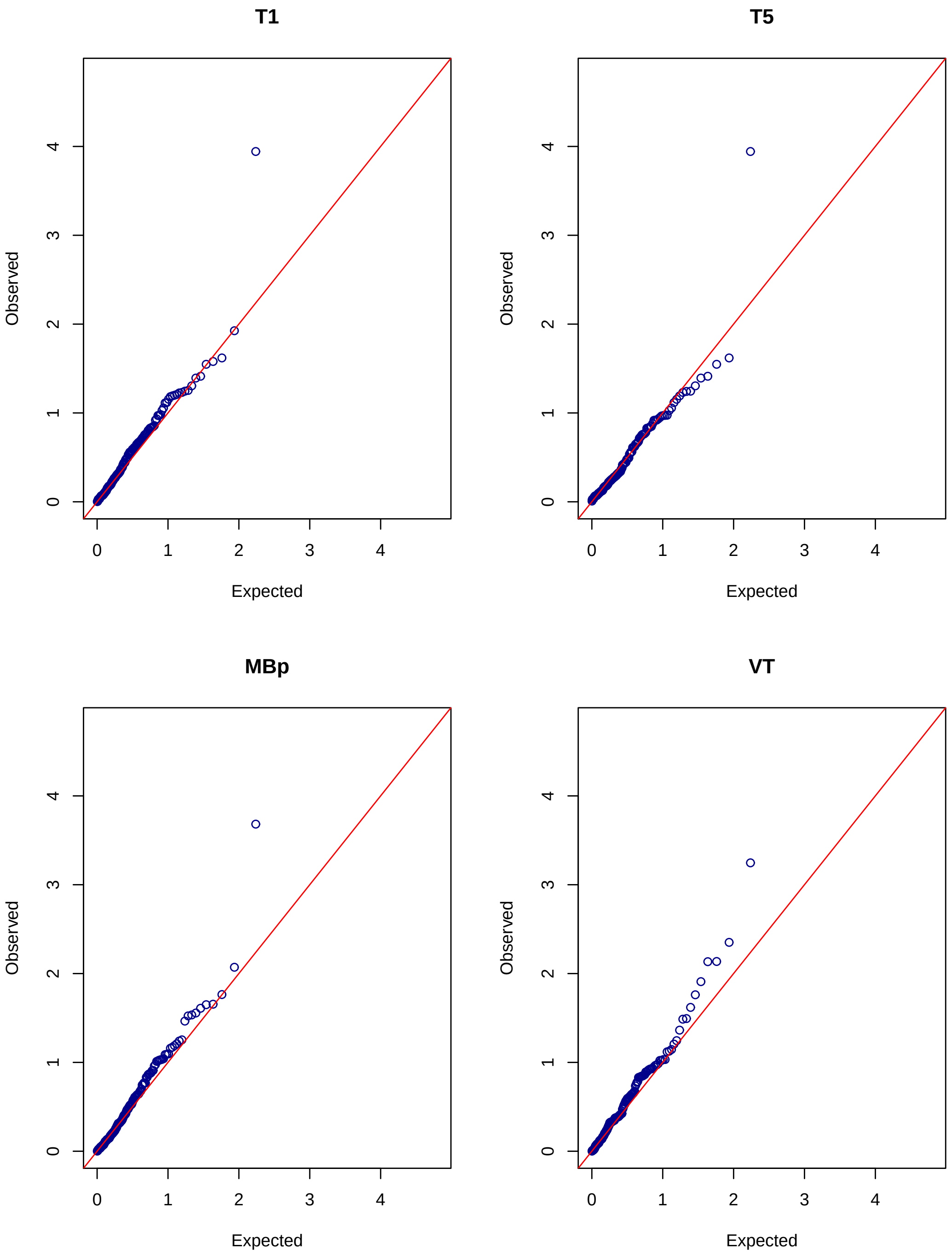


Figure 2.1: Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the asymptotic T1, T5, MB_p and VT tests in the quantitative trait analysis of total cholesterol.

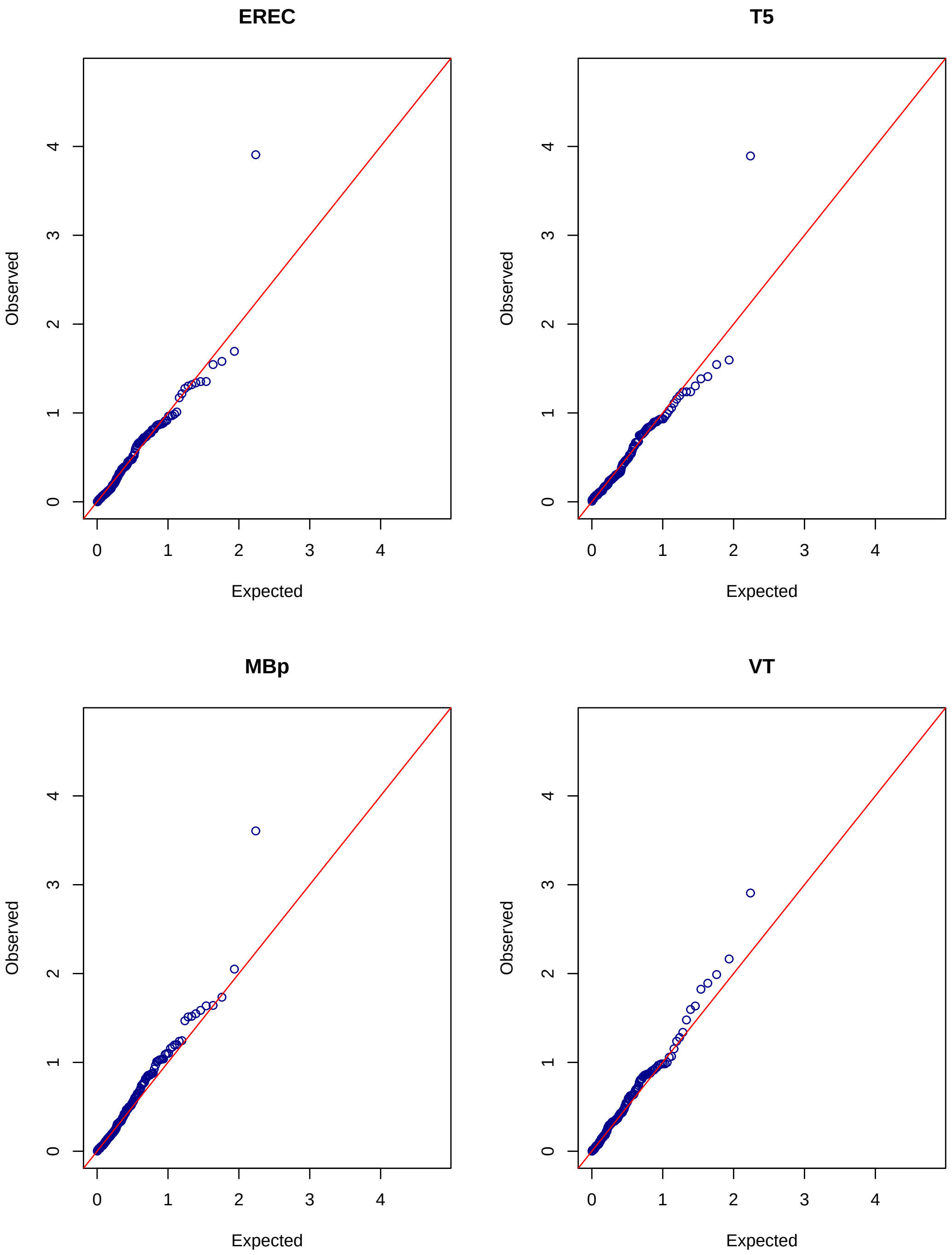


Figure 2.2: Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the permutation EREC, T5, MB_p and VT tests in the quantitative trait analysis of total cholesterol.

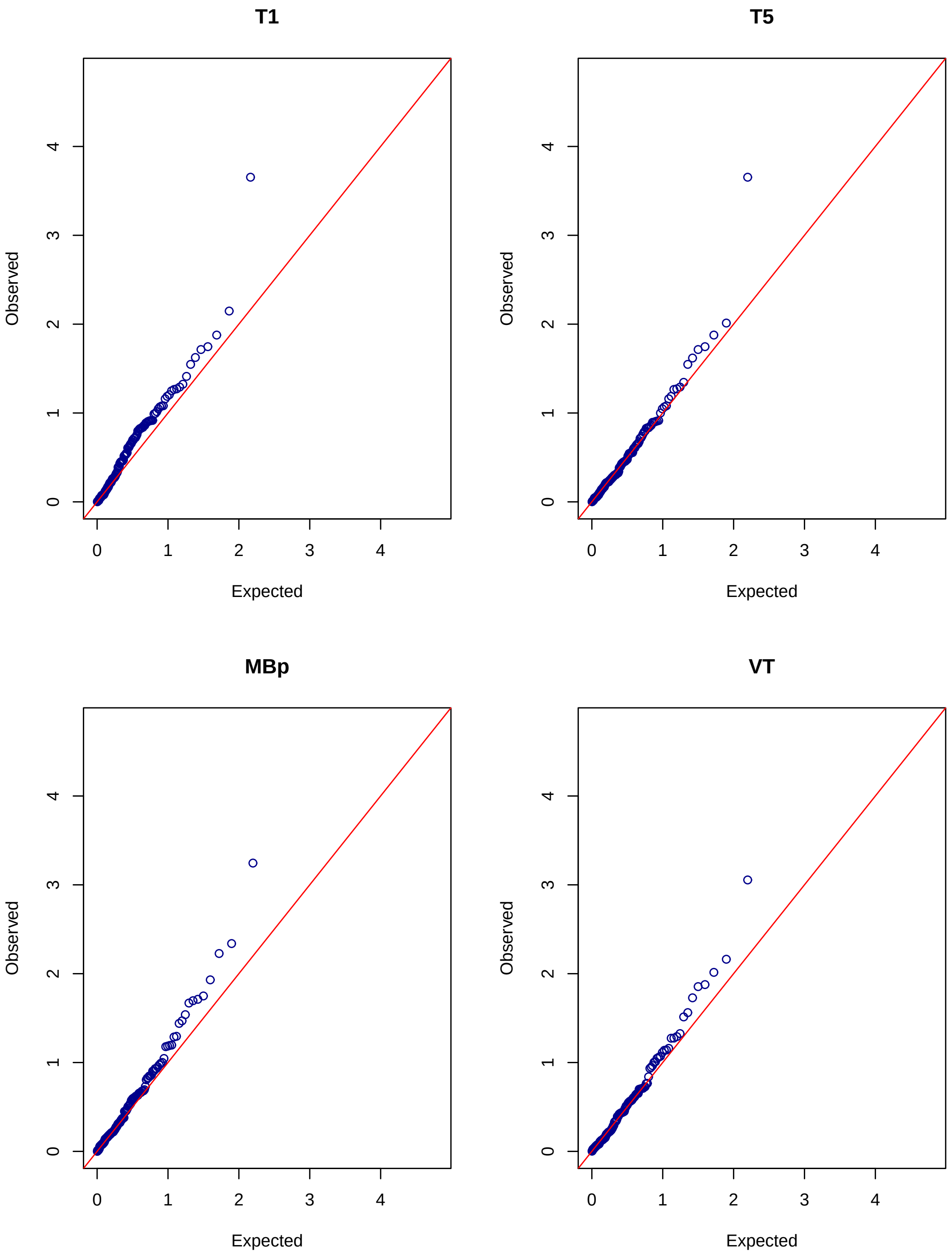


Figure 2.3: Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the asymptotic T1, T5, MB_p and VT tests in the binary trait analysis of total cholesterol.

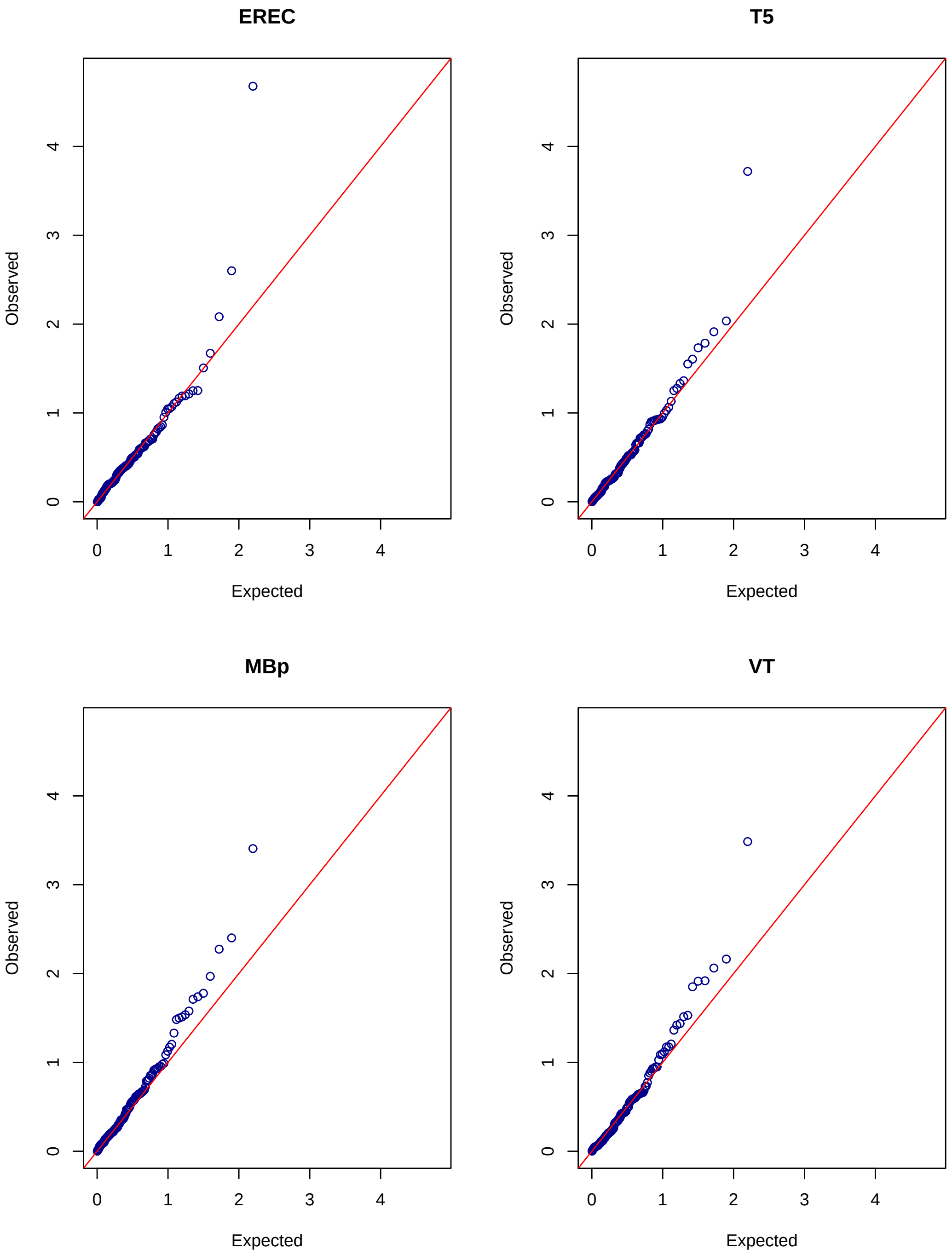


Figure 2.4: Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the bootstrap EREC, T5, MB_p and VT tests in the binary trait analysis of total cholesterol.

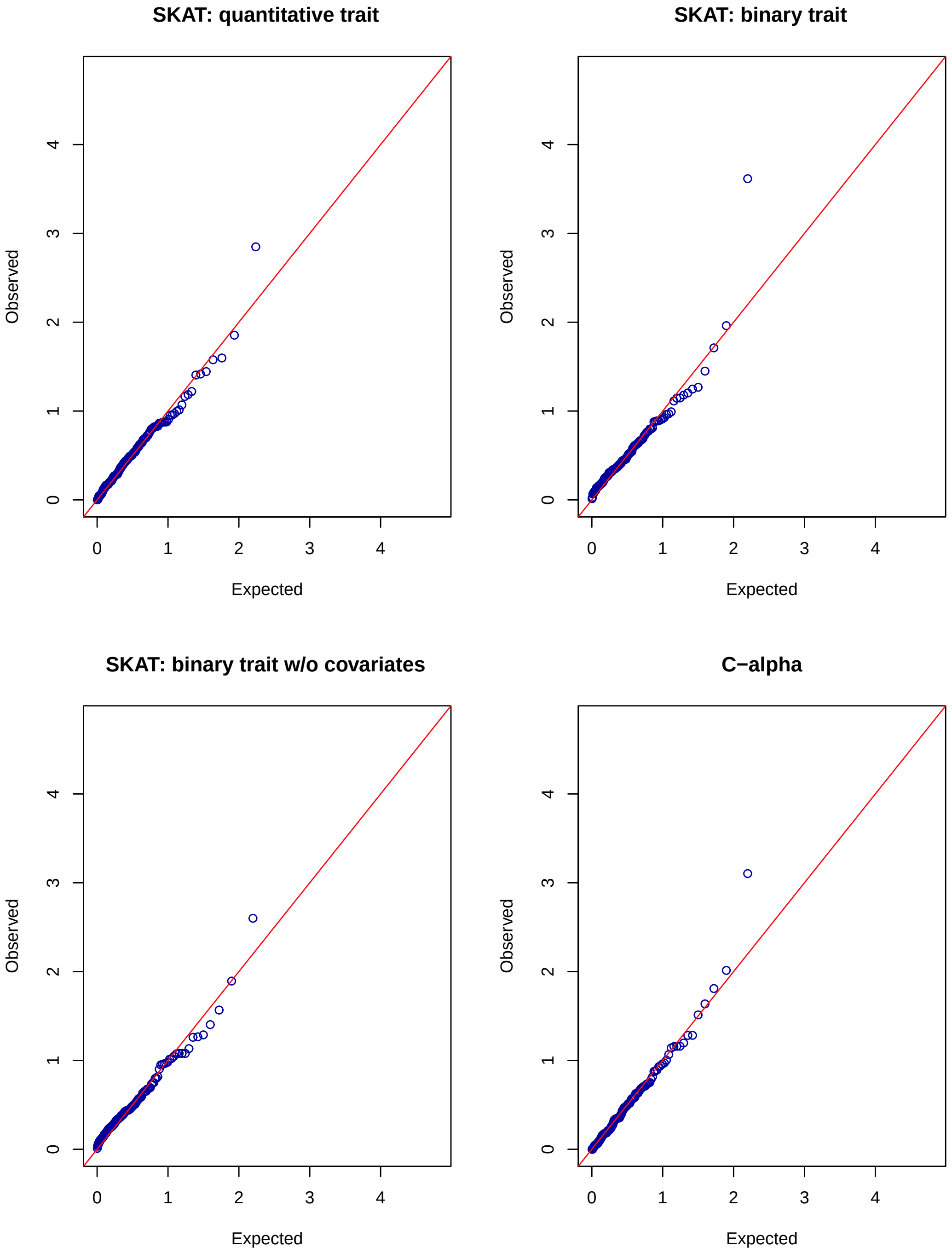


Figure 2.5: Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the SKAT in the quantitative and binary trait analyses of total cholesterol (with covariates) and for the SKAT and C-alpha in the binary trait analysis of total cholesterol without covariates.

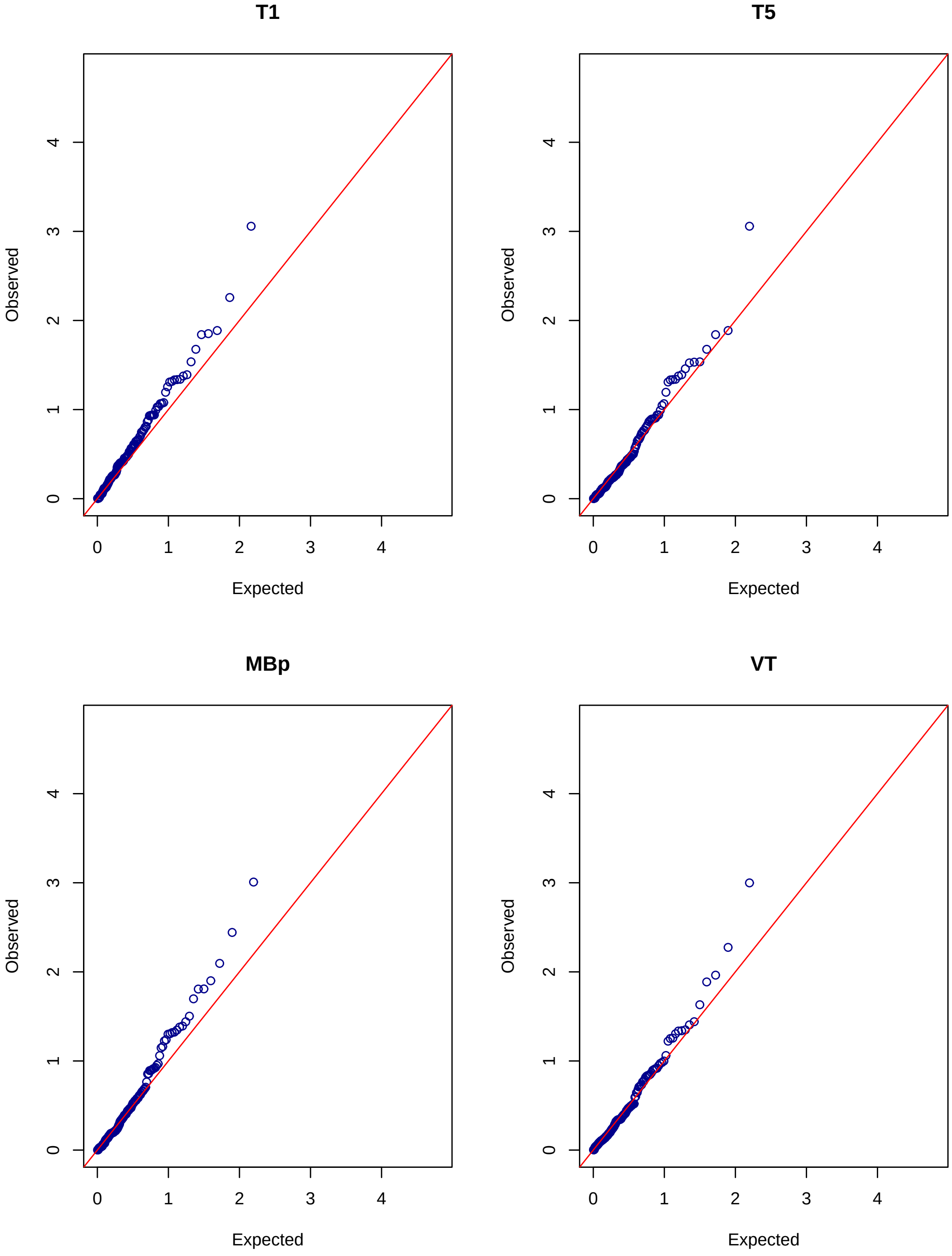


Figure 2.6: Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the asymptotic T1, T5, MB_p and VT tests in the binary trait analysis of total cholesterol without covariates.

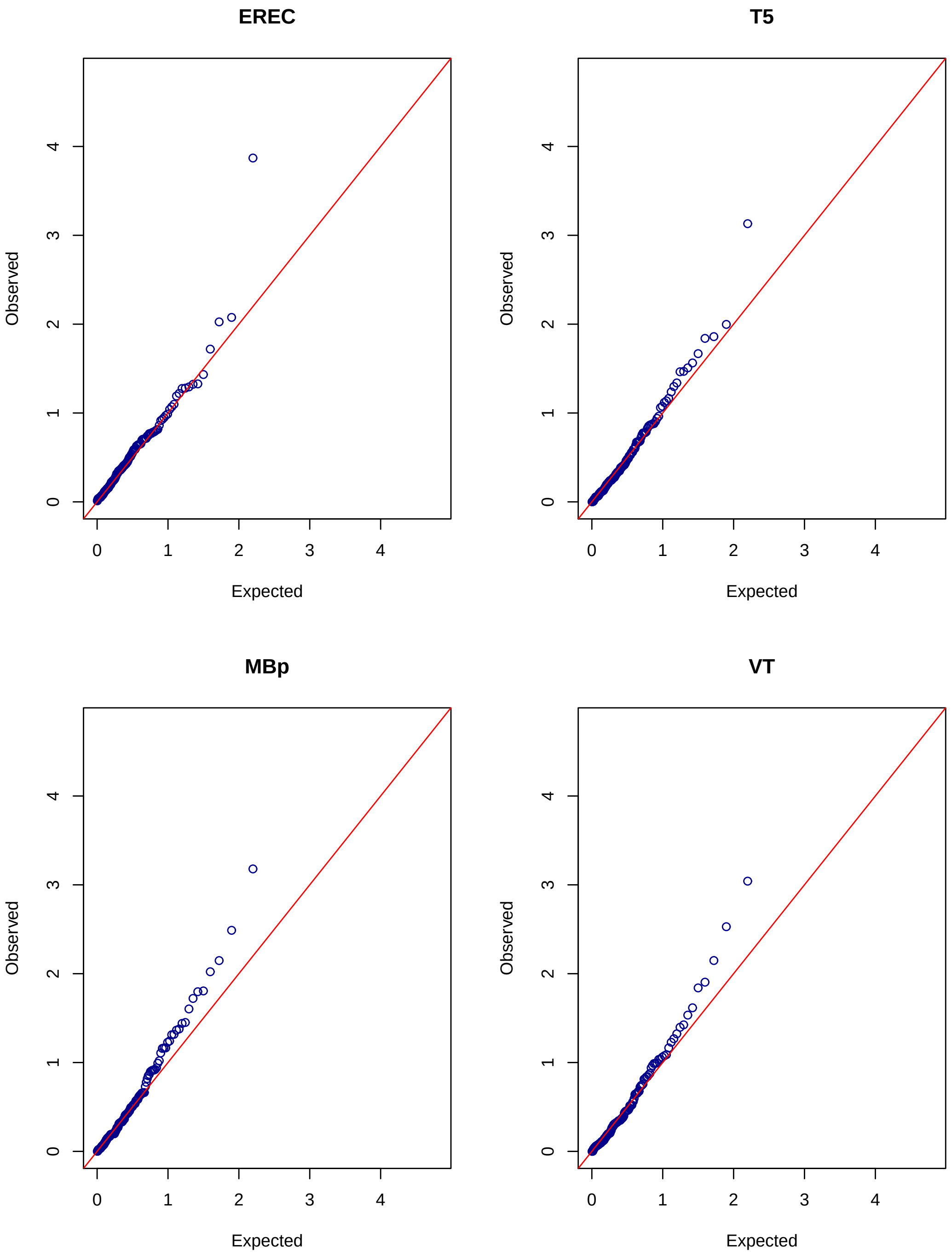


Figure 2.7: Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the permutation EREC, T5, MB_p and VT tests in the binary trait analysis of total cholesterol without covariates.

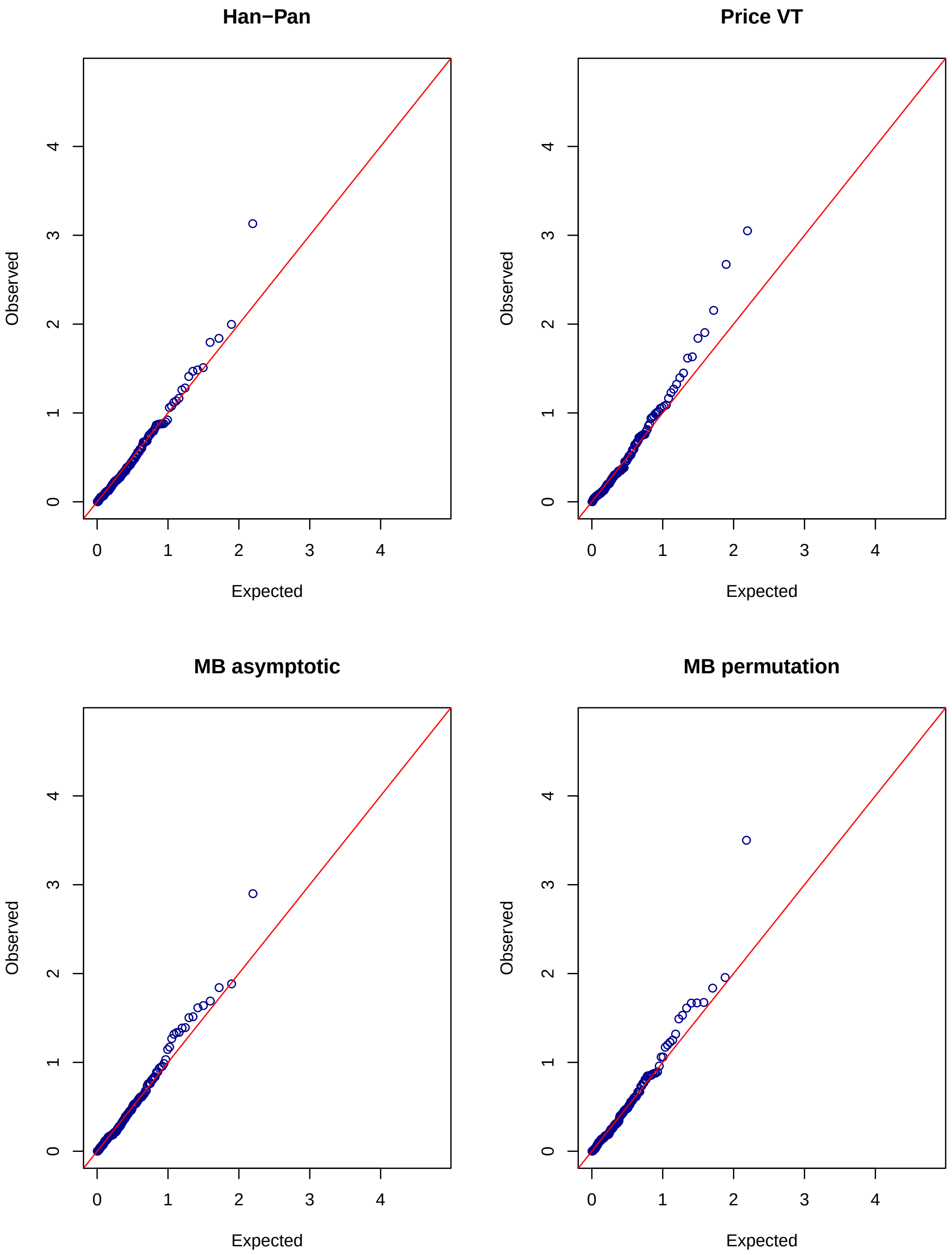


Figure 2.8: Quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the Han-Pan test, Price et al.'s VT test, and the asymptotic and permutation versions of the MB test in the binary trait analysis of total cholesterol without covariates.

2.5 Discussion

We developed a very general framework for the association analysis of rare variants. This framework enabled us to evaluate existing methods and develop new methods. Our theoretical analysis and simulation studies yielded new insights into the behavior of the existing methods. The normal approximation works very well for the new methods, and resampling is required only when the weight function depends on the phenotype values. The new methods are numerically stable and easy to implement. The asymptotic tests are extremely fast. A computer program implementing the new methods is posted at our website.

We have adopted score-type statistics, which are computationally faster and more stable than Wald and likelihood ratio (LR) statistics. Our simulation studies revealed that Wald tests tend to be overly conservative (resulting in substantial loss of power) while likelihood ratio tests tend to be too liberal (resulting in excessive false-positive findings), especially for small n and low MAFs; see Tables 2.12-2.14.

Our work improves upon the pioneer work of Madsen and Browning by using more powerful test statistics, accommodating covariates and avoiding permutation. For case-control studies, Madsen and Browning estimated the allele frequencies in the unaffected subjects only so that a true signal from an excess of mutations in the affected subjects would not be deflated by using the total number of mutations in both affected and unaffected subjects. According to our theory, the allele frequencies in the unaffected subjects will be optimal if $\log(OR_j) \propto \{p_j(1-p_j)\}^{-1/2}$ $(j = 1, \ldots, m)$ and p_j is the frequency of the jth variant in the unaffected subjects. Even if that is the truth, the frequency estimates are highly variable and can be very different from the true values. The frequency estimates in the pooled sample of affected and unaffected subjects are more stable and the corresponding MB_p-test can be implemented through normal approximation (rather than resampling).

The optimal choice of the frequency threshold depends on the nature of association, which is generally unknown. In addition, the frequency estimates for rare variants are highly variable, especially for small samples with substantial missing data. Thus, VT methods may be preferable to fixed threshold methods. Our VT approach improves upon that of Price et al. in three aspects: (1) it uses more powerful test statistics; (2) it can accommodate covariates; (3) it can be implemented by normal approximation instead of permutation.

The EREC test is capable of detecting rare mutations with opposite effects. Simulation studies (Tables 2.4 and 2.6) showed that the EREC test has similar power to the tests assuming the same direction of effects when that assumption holds and is much more powerful than the latter when that assumption fails. In addition, the EREC test outperforms the HP, C-alpha and SKAT tests. In the real data example, the EREC test produced the most convincing evidence of association for the top gene among all the tests. Thus, we recommend the EREC test for general use.

The SKAT is computationally faster than the EREC, HP and C-alpha tests because it calculates p-values analytically. Simulation studies revealed that the SKAT is overly conservative, especially when n and α are small. The resampling methods developed in this chapter can be used to obtain accurate p-values for the SKAT and indeed any other tests, with or without covariates.

It is possible to incorporate biological and computational information about the functional effects of rare variants, such as SIFT and PolyPhen scores, into the association analysis. Indeed, our theory allows incorporation of any prior knowledge into the weight function. Efficient use of functional/bioinformatics information requires further investigation. It would be worthwhile to explore Bayesian methods.

Grouping methods for rare variants are in the same vein as the SNP-set methods for GWAS studies in that multiple SNPs within a group are analyzed collectively to enhance statistical power. Because the data are extremely sparse for individual rare variants, the SNP-set methods for common variants may not be applicable to rare variants. On the other hand, the methods for rare variants can potentially be used to combine low-frequency SNPs in GWAS studies.

We have considered one group of variants at a time. It may be desirable to analyze several groups of variants simultaneously. Our approach can be readily extended to multiple groups of variants. Specifically, we divide variants into, say, K groups according to certain criteria (e.g., MAFs) and combine the information within each group. We can express the score statistic for each group of variants as a sum of n efficient score functions, so that the asymptotic joint distribution of the K score statistics follows from the multivariate central limit theorem. We can then use the asymptotic joint distribution to form a multivariate test statistic. If we choose the maximum of the K test statistics, then the formulas for K weight functions presented in the Material and Methods section can be directly applied. If we choose the chi-squared statistic with K degrees of freedom, then our method would be a generalization of the CMC of Li and Leal.

We used the Bonferroni correction in the analysis of the real data. This criterion is conservative if there is strong LD among the genes. More accurate correction for multiple testing can be achieved by accounting for the correlations of the test statistics. There are two possible ways to do so: one is to use permutation and one is to use Monte Carlo. The latter is based on efficient score functions.

CHAPTER3: BINARY SECONDARY TRAIT ANALYSIS UNDER TRAIT-DEPENDENT SAMPLING

3.1 Introduction

In this chapter, we propose a valid and efficient maximum likelihood (ML) framework for rare-variant testing of binary secondary trait associations. We model the quantitative primary trait using the approach described by Lin et al. (2013) and the binary secondary trait using a probit regression model. We evaluate statistical significance using asymptotic approximation or resampling. The resampling approach is especially important for rare-variant tests because the asymptotic approximation can be inaccurate when variants are rare and traits are binary. We compare the ML methods with the naïve methods; namely, the standard probit or logistic regression methods. We demonstrate through extensive simulations that the ML methods preserve the type I error and that the power for meta-analysis is always higher than the power for each individual study. In contrast, the naïve methods do not hold such properties under trait-dependent sampling. Finally, we apply our methods to data from NHLBI ESP.

3.2 Methods

For a given study, let Y_1 denote the quantitative primary trait and Y_2 denote a binary secondary trait. Also, let G denote the genotypes and Z denote the covariates (e.g., age, gender, principle components). In the single variant analysis, G pertains to a common variant. In the gene-level analysis, G pertains to a set of variants within a gene.

Suppose that n_1 subjects with extreme value of Y_1 are selected for sequencing from a cohort of n subjects. We assume that the primary trait Y_1 is available for all members of the cohort and the genotype G and covariate Z are available for the n_1 sequenced subjects. The secondary trait Y_2 can be available only on a subset, say n_2, of the sequenced subjects. Without loss of generality, we order the data such that the first n_2 subjects correspond to subjects with available Y_2 value and the remaining $(n_1 - n_2)$ sequenced subjects appear next. Hence, we write the observed-data likelihood in the following expression:

$$\prod_{i=1}^{n_1} P(Y_{1i} \mid G_{1i}, Z_{1i}) P(G_{1i}, Z_{1i}) \prod_{i=n_1+1}^{n} \sum_{g_1, z_1} P(Y_{1i} \mid g_1, z_1) P(g_1, z_1) \prod_{i=1}^{n_2} P(Y_{2i} | Y_{1i}, G_{2i}, Z_{2i}) \tag{3.1}$$

where G_1 and G_2 are the genetic variables (e.g., burden scores) derived from the genotype G, Z_1 and Z_2 are functions of the covariate Z, and P denotes the probability distribution function.

We postulate a continuous latent variable denoted by Y_2^* with $Y_2^* \in (-\infty, 0]$ corresponding to $Y_2 = 0$ and $Y_2^* \in (0, \infty]$ corresponding to $Y_2 = 1$. Then, we model the joint distribution of Y_1 and Y_2^* through a bivariate linear regression model:

$$Y_1 = \beta_1^{\mathrm{T}} G_1 + \gamma_1^{\mathrm{T}} Z_1 + \epsilon_1 \quad \text{and}$$

$$Y_2^* = \beta_2^{\mathrm{T}} G_2 + \gamma_2^{\mathrm{T}} Z_2 + \epsilon_2,$$

where (ϵ_1, ϵ_1) follows a bivariate normal distribution with mean 0 and covariance $\{\sigma_{ij}; i, j = 1, 2\}$. We absorb the unit component in Z such that the first components of γ_1 and γ_1 pertain to the intercepts. The distribution of Y_2^* conditional on (Y_1, G_2, Z_2) satisfies the following linear regression model

$$Y_2^* = \alpha \widetilde{Y}_1 + \beta_2^{\mathrm{T}} G_2 + \gamma_2^{\mathrm{T}} Z_2 + \widetilde{\epsilon}_2,$$

where $\alpha = \sigma_{12}/\sigma_{11}$, $\widetilde{Y}_1 = Y_1 - \beta_1^{\mathrm{T}} G_1 - \gamma_1^{\mathrm{T}} Z_1$, and $\widetilde{\epsilon}_2$ is independent of ϵ_1 and follows a normal distribution with mean 0 and variance $\widetilde{\sigma}_{22} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$. We set $\widetilde{\sigma}_{22} = 1$ because the residual variance is unidentifiable with the unobserved Y_2^*. This linear regression model is equivalent to the probit regression.

It is important to note that only the term $\prod_{i=1}^{n_2} P(Y_{2i}|Y_{1i}, G_{2i}, Z_{2i})$ in the expression 3.1 involves the parameter of interest β_2. Therefore, to make inference about β_2, we only need to focus on the probit model. However, the parameters (β_1, γ_1) in $\widetilde{Y}_1$ are unknown and need to be estimated. This can be achieved using the computationally efficient and numerically stable EM algorithm put forth by Lin et al. (2013). Let $(\widehat{\beta}_1, \widehat{\gamma}_1)$ denote the parameter estimates from the EM algorithm and $\widehat{\widetilde{Y}}_{1i}$ denote $Y_{1i} - \widehat{\beta}_1^{\mathrm{T}} G_{1i} - \widehat{\gamma}_1^{\mathrm{T}} Z_{1i}$. Then, the score statistic for the null hypothesis $H_0 : \beta_2 = 0$ can be written as

$$U = \sum_{i=1}^{n_2} \frac{(-1)^{Y_{2i}+1} \varphi\left(\widehat{\alpha}\widehat{\widetilde{Y}}_{1i} + \widehat{\gamma}_2^{\mathrm{T}} Z_{2i}\right)}{\Phi\left((-1)^{Y_{2i}+1}\left(\widehat{\alpha}\widehat{\widetilde{Y}}_{1i} + \widehat{\gamma}_2^{\mathrm{T}} Z_{2i}\right)\right)} G_{2i},$$

where $\varphi(.)$ and $\Phi(.)$ are the density function and cumulative density function respectively of the standard normal distribution, and $(\widehat{\alpha}, \widehat{\gamma}_2)$ are the restricted MLE of (α, γ_2) from the probit regression under H_0. Suppose that G_2 consists of m genetic variables $G_{2i} = (G_{21i}, \cdots, G_{2mi})$. Then, $U = (U_1, \cdots, U_m)$ is a $m \times 1$ vector and is asymptotically multivariate normal with mean 0 and covariance matrix $V = \{v_{kl}; k, l = 1, \cdots, m\}$. Various gene-level test statistics can be constructed based on U and V, in order to reflect the trait-depending sampling, and the corresponding p-value can be obtained analytically.

Alternatively, the resampling p-value can be obtained through a parametric bootstrap. Specifically, we generate the lth sample of the ith subject $Y_{2i}^{(l)}$ from the fitted null model:

$$\Pr(Y_{2i}^{(l)} = 1) = \Phi\left(\widehat{\alpha}\widehat{\widetilde{Y}}_{1i} + \widehat{\gamma}_2^{\mathrm{T}} Z_{2i}\right),$$

replace Y_{2i} with $Y_{2i}^{(l)}$, and re-calculate the test statistic. Then, the resampling p-value can be obtained by $(L_0 + 1)/(L + 1)$, where L_0 is the number of L sampled statistics that are at least as extreme as the observed statistic. For computational efficiency, we employ an adaptive procedure, in which we use small numbers of samples for the large p-values and large numbers of samples only for the small p-values.

The naïve approach is to perform standard probit or logistic regression. Two versions of the naïve method exist. The first version, referred to as probit-M or logit-M, is to regress Y_2 on (G_2, Z_2). The other version, referred to as probit-C or logit-C, is to regress Y_2 on (Y_1, G_2, Z_2). The "M" and "C" stand for "marginal" and "conditional", respectively.

3.3 Simulation studies

3.3.1 Type I error under random sampling

We investigate the type I error of the burden (aggregate of all variants), VT, and SKAT tests based on the probit and logistic regressions under random sampling, in which the binary sample can be unbalanced, especially for rare diseases. In our simulation, the data were sampled using a logistic model with one covariate. We varied the sample size and the proportion of cases. The genotypes of 10 variants were simulated, and two sets of MAFs were considered: (1) $p_j = 0.001j$ $(j = 1, \cdots, 10)$; and (2)$p_j = 0.01j$ $(j = 1, \cdots, 10)$. One million replicates were used to evaluate the type I error rate. The results are presented in Tables 3.1 and 3.2. The type I error rates for the gene-level tests based on the logistic regression model are not inflated for the unbalanced numbers of cases and controls when the MAFs of the variants are low.

Table 3.1: Type I error[a]at different nominal levels for rare-variant tests based on probit and logistic regressions under $p_j = 0.001j$ ($j = 1, \cdots, 10$). Binary data are simulated using a logistic regression model.

			Burden Test		**VT Test**		**SKAT**	
Sample Size	**Case%**	α	**Probit**	**Logit**	**Probit**	**Logit**	**Probit**	**Logit**
500	10%	10^{-2}	0.82	1.00	0.70	2.10	1.10	2.00
		10^{-3}	0.71	1.70	0.63	5.10	1.40	4.40
		10^{-4}	0.78	3.20	0.76	15.00	1.60	11.00
	20%	10^{-2}	0.91	0.98	0.68	1.40	0.80	1.20
		10^{-3}	0.80	1.20	0.45	2.10	0.73	1.70
		10^{-4}	0.68	1.50	0.36	3.20	0.54	2.60
	50%	10^{-2}	0.95	0.95	0.71	0.72	0.60	0.61
		10^{-3}	0.87	0.88	0.43	0.44	0.30	0.32
		10^{-4}	0.72	0.70	0.30	0.33	0.11	0.11
1000	10%	10^{-2}	0.92	1.00	0.74	1.90	1.10	1.60
		10^{-3}	0.86	1.30	0.67	4.60	1.50	3.10
		10^{-4}	0.80	2.40	0.81	14.00	1.70	6.60
	20%	10^{-2}	0.97	0.99	0.72	1.30	0.89	1.10
		10^{-3}	0.89	1.10	0.56	1.90	0.92	1.50
		10^{-4}	0.79	1.10	0.56	3.20	0.85	2.10
	50%	10^{-2}	0.98	0.98	0.76	0.76	0.77	0.78
		10^{-3}	0.89	0.90	0.54	0.55	0.53	0.54
		10^{-4}	0.87	0.87	0.59	0.59	0.39	0.39

[a] divided by α.

Table 3.2: Type I error[a]at different nominal levels for rare-variant tests based on probit and logistic regressions under $p_j = 0.01j$ ($j = 1, \cdots, 10$). Binary data are simulated using a logistic regression model.

			Burden Test		**VT Test**		**SKAT**	
Sample Size	**Case%**	α	**Probit**	**Logit**	**Probit**	**Logit**	**Probit**	**Logit**
500	10%	10^{-2}	0.98	1.00	0.82	1.50	0.98	1.40
		10^{-3}	0.88	1.00	0.81	2.80	1.20	2.60
		10^{-4}	0.64	1.00	0.95	7.40	1.50	5.60
	20%	10^{-2}	1.00	1.00	0.84	1.10	0.86	1.00
		10^{-3}	0.94	0.99	0.77	1.50	0.81	1.40
		10^{-4}	0.85	1.00	0.79	2.60	0.84	2.10
	50%	10^{-2}	0.99	0.99	0.89	0.89	0.81	0.81
		10^{-3}	1.00	0.99	0.75	0.76	0.65	0.66
		10^{-4}	1.00	1.00	0.76	0.79	0.44	0.44
1000	10%	10^{-2}	0.98	1.00	0.90	1.30	0.97	1.20
		10^{-3}	0.92	1.00	0.88	2.00	1.10	1.90
		10^{-4}	0.91	1.10	1.00	4.20	1.40	3.50
	20%	10^{-2}	0.99	1.00	0.92	1.10	0.92	1.00
		10^{-3}	0.97	1.00	0.87	1.30	0.91	1.20
		10^{-4}	0.95	1.00	0.90	2.00	0.90	1.60
	50%	10^{-2}	1.00	1.00	0.96	0.97	0.91	0.91
		10^{-3}	0.99	0.99	0.86	0.87	0.79	0.80
		10^{-4}	0.95	0.96	0.94	0.93	0.68	0.69

[a] divided by α.

3.3.2 Type I error and power under trait-dependent sampling

We conducted extensive simulation studies to evaluate the type I error and power of the ML methods and the naïve methods under trait-dependent sampling. Three rare-variant association tests were performed: burden test, VT test and SKAT. Population genetic data were generated using *cosi*. We simulated 10,000 haplotypes for 1000kb regions derived from the European samples. The parameters were derived from the models that best mimicked the real population (Schaffner et al. 2005). For each data set, we randomly selected 3kb genetic regions, which is the average size of the coding region of a gene (Pruitt et al. 2012). Then, we randomly paired 2 haplotypes to form the diploid of an individual. The average number of variants across the 3kb regions was about 60, of which 90% had MAF <= 5%. Because we focused on rare-variant associations, the common variants with MAF > 5% were removed.

Two quantitative traits Y_1 and Y_2^* were generated from the bivariate normal distribution:

$$\begin{pmatrix} Y_{1i} \\ Y_{2i}^* \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \beta_1^{\mathrm{T}} G + \gamma_1^{\mathrm{T}} Z \\ \beta_2^{\mathrm{T}} G + \gamma_2^{\mathrm{T}} Z \end{pmatrix}, \quad \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right),$$

where one covariate in Z was generated from the standard normal distribution. Then, the binary secondary trait Y_2 is set to 1 if Y_2^* is positive and set to 0 otherwise. We generated a cohort of 10,000 subjects and kept the values of (G, Z, Y_2) for the 250 subjects with the smallest values of Y_1 and the 250 subjects with the largest values of Y_1. In all the simulation studies, we fixed $\gamma_1 = (0, 0.2)$ and $\gamma_2 = (-1.4, 0.2)$ and varied the values of β_1 and σ_{12}. We set the nominal significance level α at 10^{-3} and obtained the p-value analytically. One million replicates were used for type I error simulations, and 10,000 replicates were used for the power simulations.

In our implementation, we let G_1 contain variants with minor alleles >= 10 and a burden score collapsing the remaining variants. For the burden test, G_2 consists of a burden score collapsing variants with MAF $< 5\%$; for the VT test, G_2 consists of burden scores based on a set of allele-frequency thresholds; and for the SKAT, G_2 consists of genotypes of individual variants.

We randomly selected 50% of the variants to be causal for the primary trait and another 50% of the variants to be causal for the secondary trait. We let all of the causal variants for the primary trait have the same positive effect sizes. The results for type I error rates are shown in Figure 3.1. Type I error rates are a little bit deflated for the ML methods under all scenarios. By contrast, type I error rates are inflated for the four naïve methods if the primary and secondary traits are correlated and if there is a genetic effect on the primary trait. The inflation for the "M" methods is more severe than the inflation for the "C" methods.

In the power simulations, we assumed that the genetic effect sizes for the secondary trait are proportional to $-log(MAF)$. In addition, positive and negative effects for the secondary trait were considered. Figure 3.2 shows the results when the effects are positive, and Figure 3.3 shows the results when the effects are negative.

3.3.3 Meta-analysis

To compare the power of the ML methods with the naïve methods for combining multiple studies, we simulated data sets for two studies and combined score statistics (Tang and Lin 2013). We note that the correlation between the primary and the secondary trait and the magnitude of the genetic effects for the primary trait are potentially different between the two studies because the primary traits in the two studies may be different. We assumed that all variants have effects on the primary trait, and we evaluated the power of the meta-analysis for the ML and the naïve methods when: (1) the correlation between the primary and secondary trait is 0.6 in study I and -0.6

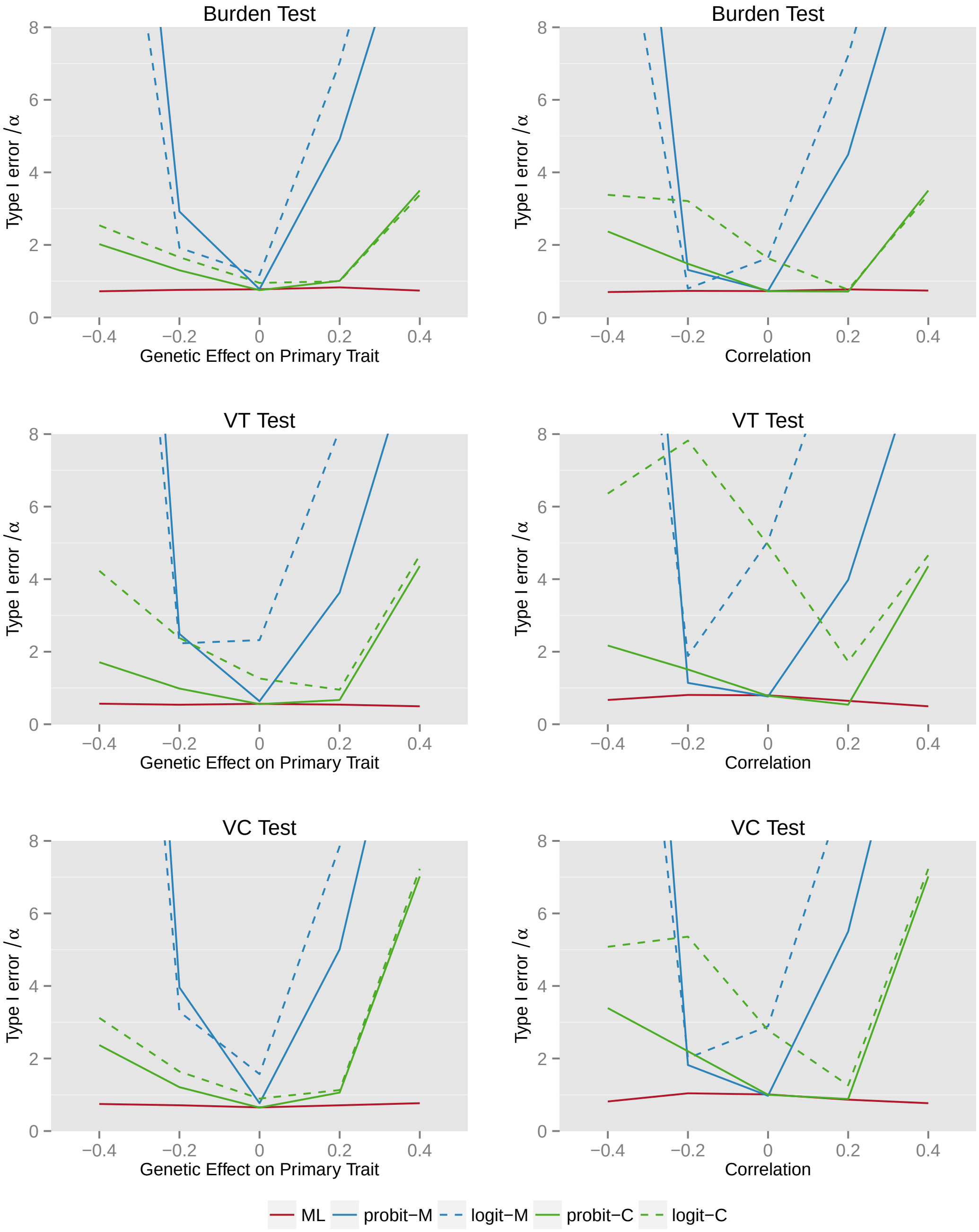


Figure 3.1: Type I error rates for the ML and naïve methods for detecting genetic associations when the genetic effects are positive. The results for the burden, VT, and VC tests are shown in the upper, middle, and lower rows, respectively. The left panel shows the power as a function of the genetic effect on the primary trait when the correlation between the primary and secondary traits is 0.4. The right panel shows the power as a function of the correlation between the primary and secondary traits when the genetic effect on the primary trait is 0.4.

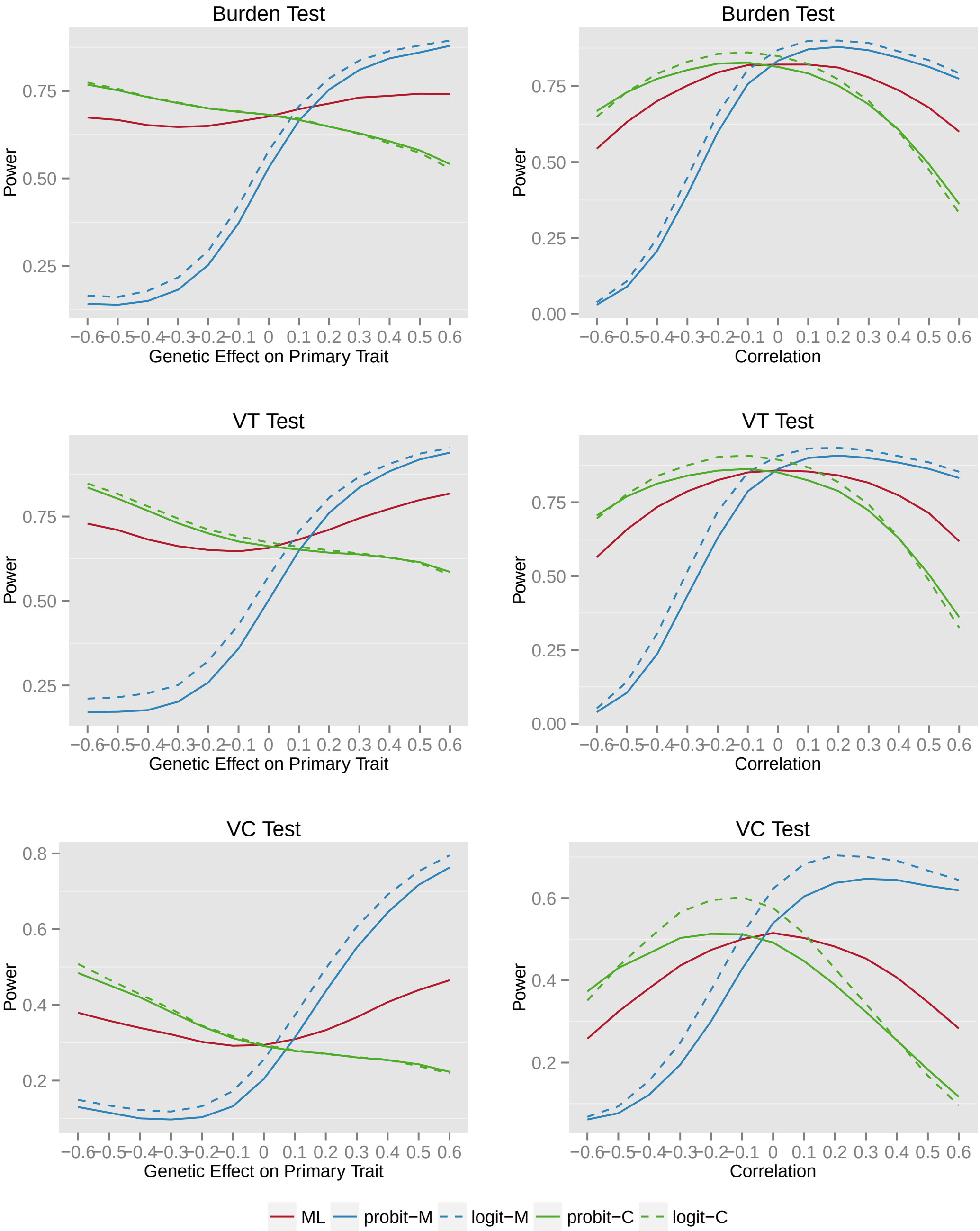


Figure 3.2: Power of the ML and naïve methods for detecting genetic associations when the genetic effects are positive. The results for burden, VT, and VC tests are shown in the upper, middle, and lower rows, respectively. The left panel shows the power as a function of the genetic effect on the primary trait when the correlation between the primary and secondary traits is 0.4. The right panel shows the power as a function of the correlation between the primary and secondary traits when the genetic effect on the primary trait is 0.4.

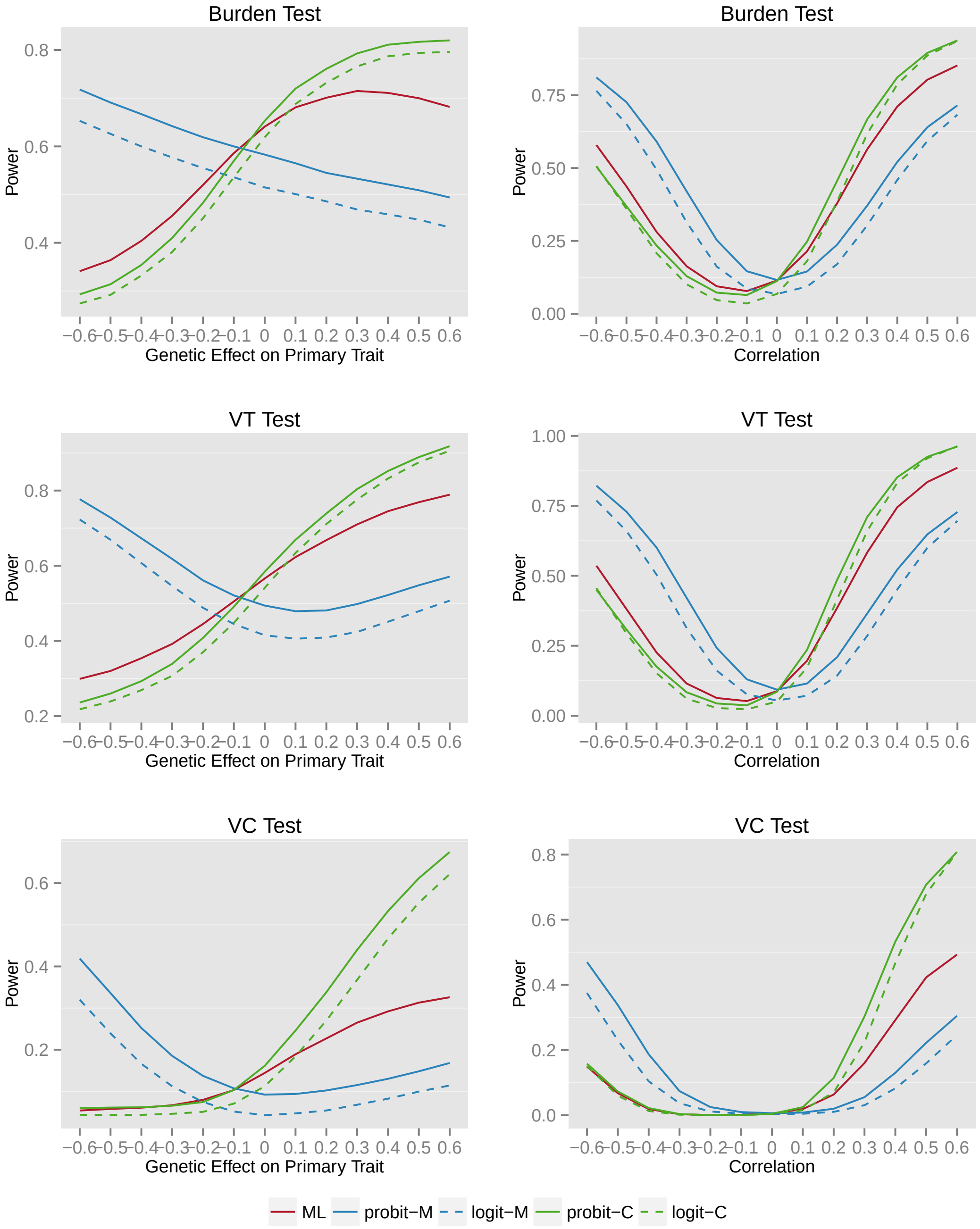


Figure 3.3: Power of the ML and naïve methods for detecting genetic associations when the genetic effects are negative. The results for burden, VT, and VC tests are shown in the upper, middle, and lower rows, respectively. The left panel shows the power as a function of the genetic effect on the primary trait when the correlation between the primary and secondary traits is 0.4. The right panel shows the power as a function of the correlation between the primary and secondary traits when the genetic effect on the primary trait is 0.4.

in study II or; (2) the genetic effect for the primary trait is 0.6 in study I and -0.6 in study II. The results are displayed in Figure 3.4. The ML methods yield larger power than the naïve methods in the meta-analysis of our simulation studies. For the ML methods, the power of the meta-analysis is always larger than the power of individual studies. This is not always the case for the naïve methods. The large power with the naïve methods for the individual studies is due to the inflation of the type I error, and the reduced power in the meta-analysis is due to the different direction of bias in the estimates for the two studies.

3.4 Data analysis

The NHLBI ESP is a signature project of the NIH Recovery Act investment: it was designed to identify genetic variants in all protein-coding regions of the human genome that are associated with heart, lung, and blood diseases. The project is comprised of multiple studies, each of which is focused on one trait. Specifically, the NHLBI ESP includes three studies that sequenced subjects with extreme values of specific quantitative trait (i.e., BMI, LDL, or BP), one case-control study on myocardial infarction (MI), and one case-only study on stroke. A total of 267 subjects with BMI $>$ 40 and 178 subjects with BMI $<$ 25 were selected for sequencing out of a total of 11,468 subjects from one cohort. For the LDL study, 120 subjects with the highest LDL values and 120 subjects with the lowest LDL values were selected out of ~22,000 European Americans from 4 cohorts; likewise, 120 subjects with the highest LDL values and 120 subjects with the lowest LDL values were selected out of ~7,000 African Americans from the same cohorts. The LDL values were adjusted for age, sex, and lipid medication, and the adjusted LDL values for the selected subjects represented less than the first percentile and greater than the 99th percentile for European ancestry and less than the 3rd percentile and greater than the 97th percentile for African ancestry. For

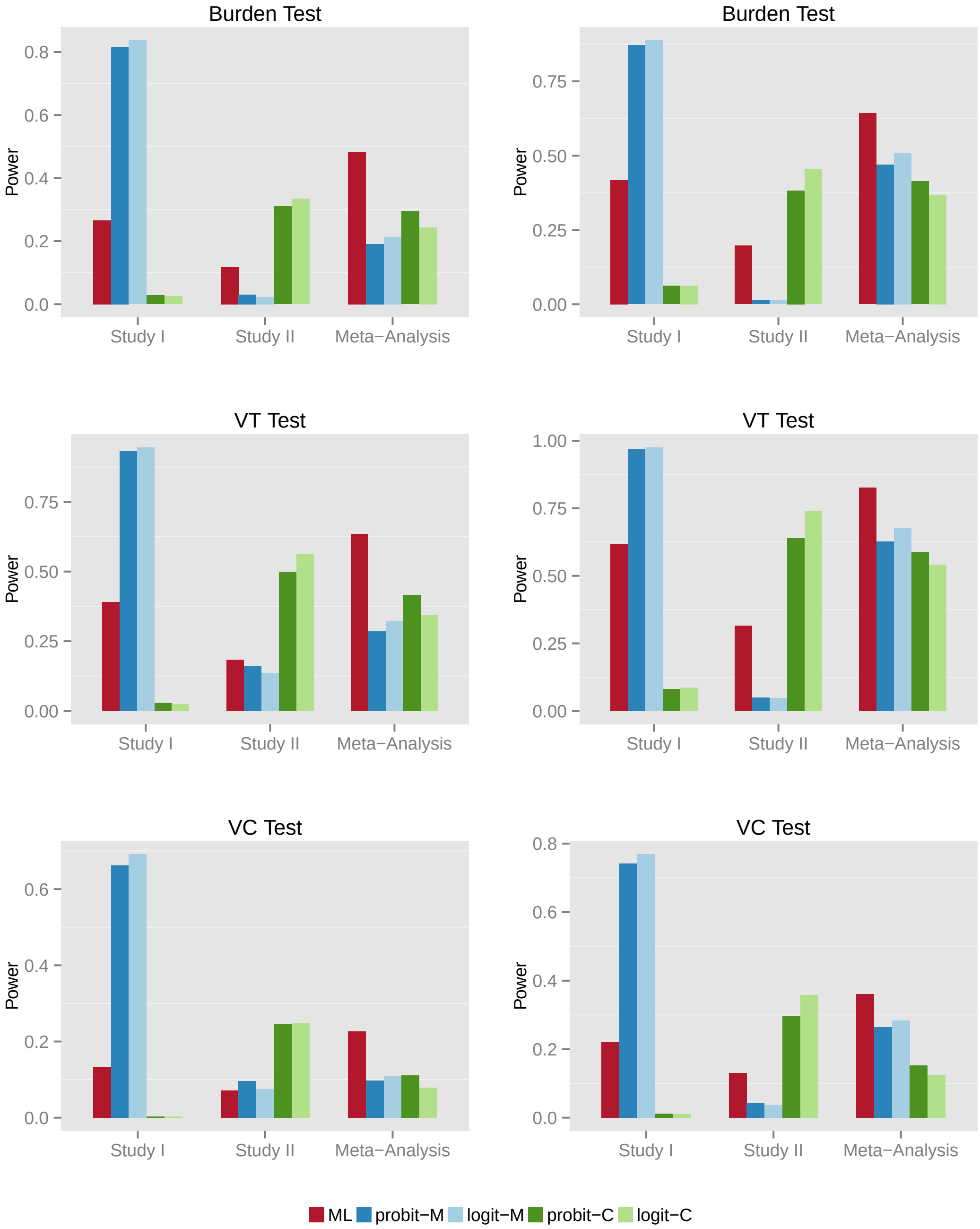


Figure 3.4: Power of the meta-analysis using the ML and naïve methods for detecting genetic associations. The results for the burden, VT, and VC tests are shown in the upper, middle, and lower rows, respectively. For the left panel, we set the genetic effect on the primary trait in both studies to 0.4 and the correlations between the primary trait and the secondary trait in Study I and Study II to 0.6 and -0.6, respectively. For the right panel, we set the correlations between the primary trait and the secondary trait in both studies to 0.4 and the genetic effect on the primary trait in Study I and Study II to 0.6 and -0.6, respectively.

the BP study, 850 subjects were selected from the top and bottom 0.2% to 1.0% of the BP distribution (adjusted for sex, age, race, BMI, and antihypertensive treatment) for ∼100,000 European Americans and ∼20,000 African Americans from 7 cohorts. The MI case-control study was comprised of 650 cases with early MI and 650 controls free of MI. The stroke case-only study was comprised of of 600 subjects with ischemic stroke. In addition to the above five studies, the NHLBI ESP included a random sample of 1,000 European Americans and 500 African Americans who had a predefined common set of core variables (i.e., phenotypes and traits); this cohort is referred to as the deeply phenotyped reference (DPR). Whole-exome sequencing was performed at the University of Washington and the Broad Institute.

For illustration, we considered the status of diabetes as the trait of interest, which is the secondary trait in the BMI, LDL, BP, MI, and stroke studies. We used both the ML and naïve methods to analyze associations with diabetes in the LDL, BMI, and BP studies and performed standard logistic regression analysis to identify associations with diabetes in the MI study (adjusted for MI status), stroke study, and DPR. (We note that both early MI and ischemic stroke are relatively rare. For a case-control or case-only study on a rare disease, standard logistic regression analysis of secondary quantitative traits conditional on the disease status yields approximately correct results.) After restricting our analysis to subjects with available sequencing data and excluding subjects with sex-mismatch or relatedness, there were 627, 632, 766, 1634, 499, and 950 subjects in the LDL, BMI, BP, MI, stroke, and DPR studies, respectively. After further restricting to subjects with available diabetes status, there were 607, 628, 693, 1553, 498, and 950. To ensure high-quality genotype calls for the analysis, we set the individual genotype values to missing if the genotype depth (GD) was lower than 10. To reduce the number of variant calls resulting from sequencing and alignment artifacts, we adopted a support vector machine (SVM) to separate likely true positives

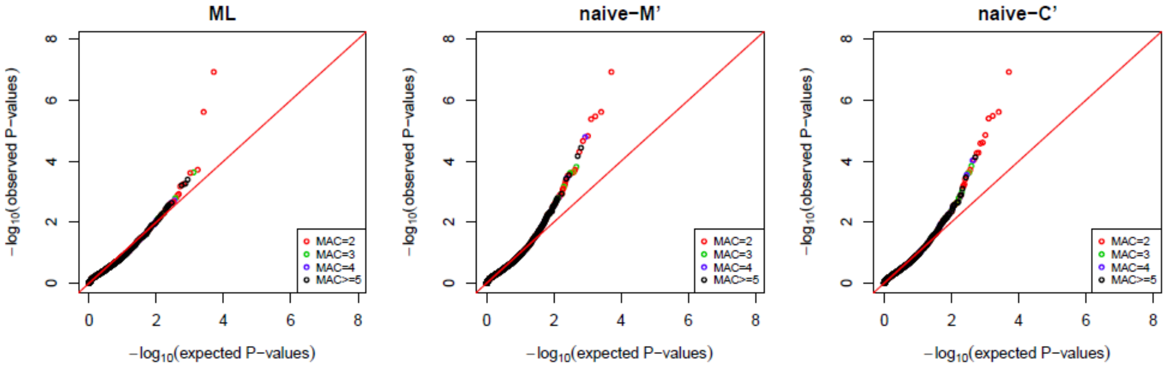


Figure 3.5: Quantile-quantile plots for the ML, naïve-M', and naïve-C' meta-analysis results.

from likely artifacts by using a variety of SNP quality metrics, including allelic balance (i.e., the proportional representation of each allele in likely heterozygotes), base quality distribution for sites supporting the reference and alternate alleles, and the distribution of supporting evidence between strands and sequencing cycle. After these quality control filters, there were a total of 115,515 SNPs with call rates $> 90\%$ and MAFs $\geq 0.5\%$. Approximately 60% of the SNPs were in exons. We focused on single-variant analysis under the additive mode of inheritance and included the top five principal components for ancestry, age, squared age, gender, race, cohort, and sequencing center/target as covariates. The natural logarithm was applied to the LDL and BMI values.

Figure 3.5 displays the quantile-quantile plots for the meta-analysis of the six studies based on ML, naïve-M' (logit-M') and naïve-C'(logit-C'), and Figure 3.6 shows the pair plots between the p-values of the ML method and the naïve methods. For the MLE meta-analysis, the observed p-values agree very well with the global null hypothesis of no association except at the extreme right tails. By contrast, the observed p-values of Naïve-M' and Naïve-C' deviate substantially from the null distribution, reflecting excessive false positive results.

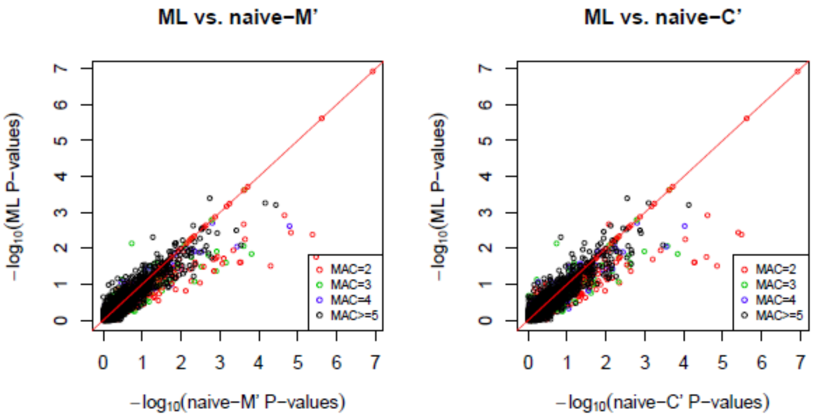


Figure 3.6: Pair plots between the p-values for the ML method and the naïve methods.

CHAPTER4: META-ANALYSIS IN SEQUENCING STUDIES

4.1 Introduction

Meta-analysis, which combines summary statistics from multiple studies, is an important tool to boost statistical power. Several research groups have recently developed meta-analysis methods for gene-level association tests (Tang and Lin 2013, Lee et al. 2013, Liu et al. 2014, Hu et al. 2013). Most of those methods assume fixed-effects (FE) models, under which the genetic effects are the same in all participating studies. If the populations or environmental factors differ substantially among studies, then the effect sizes will likely be unequal, especially for rare variants. This phenomenon is referred to as (between-study) heterogeneity, which may also be caused by different definitions or measurements of the phenotype and different collections or manipulations of genotype data (e.g., sequencing platforms and quality control criteria) (Ioannidis et al. 2007, Waters et al. 2010, Heid et al. 2009; 2010, Tobacco and Consortium 2010). Further heterogeneity may arise from differences in gene annotation, variant selection and MAF calculation.

In this chapter, we propose simple meta-analysis methods for gene-level association tests under random-effects (RE) models, which allow the genetic effects to vary among studies. Our methods take score statistics, rather than individual participant data, as input and thus can accommodate any study designs (e.g., case-control, cross-sectional, cohort, and family studies) and any phenotypes (e.g., binary, quantitative, and censored). We produce the RE versions of all commonly used gene-level association tests, including burden, CMC, VT, VC and SKAT-O. Each test statistic provides a joint test of the mean and the variation of the genetic effects among the studies and thus has high power when the average effect is large or the heterogeneity is strong or both. We demonstrate through extensive simulation studies that our RE methods are substantially more powerful than the FE methods in the presence of moderate and high heterogeneity and are nearly as powerful as the latter when the heterogeneity is low. We illustrate the usefulness of the proposed methods through an application to the NHLBI ESP.

4.2 Methods

Suppose that we are interested in the effects of d genetic variables on a particular phenotype. For the burden test, the genetic variable is the burden score. For the CMC test, the genetic variables consist of the burden scores for rare variants and the genotypes for common variants. For the VT test, the genetic variables are the burden scores at the observed MAF thresholds. For the VC test, the genetic variables are the genotypes of individual variants.

We wish to perform meta-analysis of K independent studies. For $k = 1, \ldots, K$, let $\beta_k = (\beta_{k1}, \cdots, \beta_{kd})^{\mathrm{T}}$ denote the effects of the d genetic variables in the kth study. It is natural to postulate the following random-effects model:

$$\beta_k = \mu + \xi_k, \quad k = 1, \ldots, K, \tag{4.1}$$

where $\mu = (\mu_1, \cdots, \mu_d)^{\mathrm{T}}$ represents the average genetic effects among the studies, and $\xi_k = (\xi_{k1}, \cdots, \xi_{kd})^{\mathrm{T}}$ is a set of random effects representing the deviations of the genetic effects of the kth study from the average effects. It is assumed that ξ_k follows a multivariate normal distribution with mean 0 and covariance matrix Σ.

We are interested in testing the null hypothesis that the d genetic variables are not associated with the phenotype in any of the K studies, i.e., $\beta_1 = \beta_2 = \cdots = \beta_K = 0$. This null hypothesis corresponds to $H_0 : \mu = 0$ and $\Sigma = 0$ under model 4.1. When the dimension d is large, the statistic for testing H_0 with an arbitrary Σ will have many degrees of freedom and thus have limited power. To increase power, we impose some structure on Σ by writing $\Sigma = \sigma B$, where σ is an unknown constant, and B is a prespecified matrix. Since $\sigma = 0$ is equivalent to $\Sigma = 0$, the null hypothesis H_0 can be written as $H_0 : \mu = 0$ and $\sigma = 0$.

In practice, the true structure of B is unknown. It is reasonable to assume compound symmetry such that

$$B = \begin{bmatrix} b_1^2 & b_1 b_2 r & \ldots & b_1 b_d r \\ b_2 b_1 r & b_2^2 & \ldots & b_2 b_d r \\ \vdots & \vdots & \ddots & \vdots \\ b_d b_1 r & b_d b_2 r & \cdots & b_d^2 \end{bmatrix},$$

where $(b_1, \cdots b_d)$ controls the relative degrees of heterogeneity for the d genetic effects, and r specifies the correlation of heterogeneity. If we believe that heterogeneity is higher for rarer variants, then we let the b_j's be inversely related to the MAFs. If the variations of the d effects are independent, then $r = 0$. In constructing the test statistics, we may set r to a certain value, say 0, or vary r from 0 to 1. It is important to point out that the choice of B affects the power but not the type I error since $\sigma = 0$ entails $\Sigma = 0$ regardless of the value of B. As will be seen later, B is involved only in the CMC and VC tests.

For the kth study, we obtain the d-dimensional score statistic U_k for testing the null hypothesis that $\beta_k = 0$ and the corresponding information matrix V_k. We describe below how to use the U_k's and V_k's to construct the RE versions of the burden, VT, VC and related tests. The derivations are given in the Appendix.

For the simple burden test, there is only one genetic variable, which is the burden score. The score statistic for testing the null hypothesis $H_0 : \mu = 0$ and $\sigma = 0$ is

$$\text{RE-BS} = \Big(\sum_{k=1}^{K} U_k\Big)^2 \Big/ \sum_{k=1}^{K} V_k + \frac{1}{2}\Big(\sum_{k=1}^{K} U_k^2 - \sum_{k=1}^{K} V_k\Big)^2 \Big/ \sum_{k=1}^{K} V_k^2. \tag{4.2}$$

The first term, denoted by FE-BS, pertains to the score statistic for testing $\mu = 0$ under the fixed-effects model ($\sigma = 0$) and the second term to the score statistic for testing $\sigma = 0$ given $\mu = 0$. The two statistics are combined through direct summation because they are uncorrelated. Since it is a joint test of the mean and heterogeneity of the effects, RE-BS will have high power when the mean effect size is large or/and when the between-study heterogeneity is strong.

For the CMC (Li and Leal 2008) and other tests involving multiple burden scores, the test statistic takes a multivariate form

$$\text{RE-CMC} = U_\mu^{\mathrm{T}} V_\mu^{-1} U_\mu + \frac{U_\sigma^2}{V_\sigma}, \tag{4.3}$$

where $U_\mu = \sum_{k=1}^{K} U_k$, $V_\mu = \sum_{k=1}^{K} V_k$, $U_\sigma = \frac{1}{2}\sum_{k=1}^{K} U_k^{\mathrm{T}} B U_k - \frac{1}{2}\mathrm{tr}(V_\mu B)$,
$V_\sigma = \frac{1}{2}\mathrm{tr}\big(\sum_{k=1}^{K} V_k B V_k B\big)$, and tr stands for trace. If $d = 1$, then 4.3 reduces to 4.2. When $d > 1$, we set $r = 0$. Alternatively, we may choose the value of r that yields the smallest p-value for RE-CMC. The resulting test statistic is denoted by RE-CMC-O, where O means that the test statistic is "optimized" over r. The calculation of the p-value for RE-CMC-O needs to account for the fact multiple values of r have been tried.

The asymptotic approximations to the distributions of RE-BS, RE-CMC and RE-CMC-O require large K and may not be accurate for small K. Thus, we use Monte Carlo simulation to obtain the p-values for these tests and all subsequent ones. To be specific, we repeatedly generate U_k from the d-variate normal distribution with mean

0 and covariance matrix V_k for $k = 1, \cdots, K$ and recalculate the test statistic. The p-value is set to be the proportion of the simulated test statistics that are greater than the observed test statistic. To improve computational efficiency, we employ an adaptive procedure which uses a small number of simulations for a large p-value and a large number of simulations for an extreme p-value. Specifically, we use 1,000 simulations for p-values greater than 0.1, 100,000 simulations for p-values between 0.001 and 0.1, 1 million simulations for p-values less than 0.001. This adaptive strategy makes Monte Carlo simulation almost as fast as the asymptotic approximation since most genes have large p-values.

For the VT method, the genetic variables correspond to the burden scores at d MAF thresholds. We perform a burden test at each MAF threshold and choose the threshold that produces the largest test statistic. Thus, the VT test statistic is defined by

$$\text{RE-VT} = \max_{j=1,\ldots,d} \left\{ u_j^2/v_j + \frac{1}{2}\left(\sum_{k=1}^{K} u_{kj}^2 - \sum_{k=1}^{K} v_{kj} \right)^2 \Big/ \sum_{k=1}^{K} v_{kj}^2 \right\},$$

where u_j and u_{kj} are the jth components of U_μ and U_k, respectively, and v_j and v_{kj} are the jth diagonal elements of V_μ and V_k, respectively. The FE counterpart is FE-VT $= \max_{j=1,\ldots,d} u_j^2/v_j$.

For the VC test, the genetic variables consist of the individual genotypes of d variants. We assume that the set of average genetic effects μ is a d-variate normal random vector with mean 0 and covariance matrix τW, where τ is an unknown constant, and W is a pre-specified matrix. We impose compound symmetry such that

$$W = \begin{bmatrix} w_1^2 & w_1 w_2 \rho & \ldots & w_1 w_d \rho \\ w_2 w_1 \rho & w_2^2 & \ldots & w_2 w_d \rho \\ \vdots & \vdots & \ddots & \vdots \\ w_d w_1 \rho & w_d w_2 \rho & \cdots & w_d^2 \end{bmatrix},$$

where $(w_1, \cdots w_d)$ controls the relative magnitutes of the d average genetic effects, and ρ indicates the correlation of the d effects. Note that W measures the within-study random effects of individual variants whereas B measures the between-study heterogeneity.

Since $\tau = 0$ is equivalent to $\mu = 0$, the null hypothesis H_0 becomes $\tau = \sigma = 0$. The score statistic for testing H_0 takes the form

$$\text{RE-VC} = \begin{bmatrix} U_\tau & U_\sigma \end{bmatrix} V_{\tau\sigma}^{-1} \begin{bmatrix} U_\tau \\ U_\sigma \end{bmatrix},$$

where $U_\tau = \frac{1}{2} U_\mu^{\mathrm{T}} W U_\mu - \frac{1}{2} \mathrm{tr}(V_\mu W)$, U_μ, V_μ, and U_σ were defined below equation 4.3 but now pertain to individual variants instead of burden scores, and

$$V_{\tau\sigma} = \frac{1}{2} \begin{bmatrix} \mathrm{tr}\left(V_\mu W V_\mu W\right) & \mathrm{tr}\left(\sum_{k=1}^{K} V_k W V_k B\right) \\ \mathrm{tr}\left(\sum_{k=1}^{K} V_k B V_k W\right) & \mathrm{tr}\left(\sum_{k=1}^{K} V_k B V_k B\right) \end{bmatrix},$$

which is the covariance matrix of (U_τ, U_σ). The FE version is FE-VC $= 2U_\tau^2 / \mathrm{tr}\left(V_\mu W V_\mu W\right)$. As in the case of RE-BS, RE-CMC and RE-VT, both the mean and heterogeneity contributes to RE-VC; however, the two contributions are correlated and thus cannot be directly added.

In original VC tests, ρ is set to 0 to allow the multiple effects within a gene to vary independently. By default, we set $\rho = 0$ for FE-VC and $\rho = r = 0$ for RE-VC. If $\rho = r = 1$, then RE-VC would become RE-BS. We can choose the value of ρ that yields the smallest p-value for FE-VC and the combination of ρ and r that yields the smallest p-value for RE-VC. The resulting test statistics are denoted by FE-VC-O and RE-VC-O, respectively. FE-VC-O is a standardized version of SKAT-O, and RE-VC-O can be viewed as a RE version of SKAT-O.

RE-BS is optimal if the effects of individual variants are similar within each study.

RE-VT allows the choice of the MAF threshold to be data-dependent. RE-VC is desirable if the effects of individual variants within a study are different. RE-VC-O allows the data to suggest how the effects of individual variants vary within and between studies.

4.3 Simulation studies

We conducted extensive simulation studies to evaluate the performance of the proposed and existing methods. We considered meta-analysis of five studies with sample sizes of 800, 1,000, 1,200, 1,400, and 1,600. Following Liu et al. (2014), we generated 12,000 haplotypes of length 1000kb under a calibrated coalescent model (Hudson 2002) mimicking a sample of three European populations (Kryukov et al. 2009). The model includes an ancient bottleneck, recent exponential growth, differentiation and migration. For each simulated data set, we randomly selected ten 300 base-pair regions to construct a 3kb region, which is the average size of the coding region of a gene (Pruitt et al. 2012). The MAFs were $< 1\%$ for 97% of the polymorphic sites. We removed variants with MAFs>5%.

We considered both quantitative and binary traits. For the quantitative trait, we generated data from the linear regression model

$$Y_{ki} = \beta_k^{\mathrm{T}} G_{ki} + \gamma_k^{\mathrm{T}} Z_{ki} + \epsilon_{ki},$$

where G_{ki} consists of the genotypes of the variants in the gene for the ith subject of the kth study, Z_{ki} consists of 1 and a normal random variable with unit variance and with mean being the total minor allele count for the ith subject of the kth study, and ϵ_{ki} is standard normal. The normal covariate represents a principal component for ancestry or a different genetically related variable. For the binary trait, we generated case-control data with an equal number of cases and controls from the logistic regression model

$$\mathrm{logit} P(Y_{ki} = 1) = \beta_k^{\mathrm{T}} G_{ki} + \gamma_k^{\mathrm{T}} Z_{ki}.$$

We set the intercepts in the linear and logistic regression models to 0 and -2, respectively, and set the regression coefficients for the normal covariate to 0.3. We compared ten meta-analysis methods: FE-BS, FE-VT, FE-VC and FE-VC-O pertain to fixed-effects models; RE-BS, RE-VT, RE-VC and RE-VC-O are our proposed methods under random-effects models; Het-SKAT and Het-SKAT-O are Lee et al. (2013)'s tests for heterogeneous effects. For the burden tests (FE-BS and RE-BS), the burden score was a weighted sum of the mutation counts with the jth variant receiving the weight $\mathrm{Beta}(\mathrm{MAF}_j; 1, 25)$, where MAF_j is the MAF of the jth variant estimated from all study subjects. (The beta function gives more weights to rarer variants.) We set the w_j's and b_j's involved in the VC tests (FE-VC, RE-VC, FE-VC-O and RE-VC-O) according to $\mathrm{Beta}(\mathrm{MAF}_j; 1, 25)$. For FE-VC-O, we did a grid search over $\rho = (0, 0.5, 1)$. For RE-VC-O, we added a grid search over $r = (0, 0.5, 1)$. We implemented Het-SKAT and Het-SKAT-O via the MetaSKAT software (Lee et al. 2013).

We used 1 million replicates to evaluate the type I error at the nominal significance level $\alpha = 10^{-2}$, 10^{-3} and 10^{-4} by setting $\beta_1 = \beta_2 = \ldots = \beta_5 = 0$. The results are shown in Table 4.1. All our tests have accurate control of the type I error, although the RE-VT test appears to be slightly conservative for the binary trait. Het-SKAT and Het-SKAT-O tend to be conservative for both the quantitative and binary traits.

We used 10,000 replicates to evaluate the power at $\alpha = 10^{-4}$. In each replicate, we randomly selected 80%, 50% or 20% of the variants to be potentially causal. Let m denote the total number of potentially causal variants. We determined the genetic effects $\beta_k = (\beta_{k1}, \ldots, \beta_{km})^{\mathrm{T}}$ by specifying the average effects $\mu = (\mu_1, \ldots, \mu_m)^{\mathrm{T}}$ and the random effects $\xi_k = (\xi_{k1}, \ldots, \xi_{km})^{\mathrm{T}}$ in model 4.1. The genetic effects were allowed to exhibit at the burden score or individual variant level. Because rarer variants tend to have larger effects on complex diseases (Pritchard 2001, Gorlov et al. 2008), we set the effect sizes of the m variants according to their MAFs through a beta function. Specifically, we generated three different structures of genetic effects: (a) set $\mu_j = a_j\theta$ and $\xi_{kj} = a_j\delta_k$ ($j = 1, \ldots, m$), where a_j is given by the Beta($\text{MAF}_j; 1, 25$) function, θ is a constant, and δ_k is a normal random variable with mean 0 and variance σ; (b) set μ_j and ξ_{kj} to be the same as under structure (a) if $\text{MAF}_j < 1\%$ and set $\mu_j = \xi_{kj} = 0$ otherwise; (c) set μ_j to be a normal random variable with mean 0 and variance $a_j\tau$, and set ξ_{kj} to be a normal random variable with mean 0 and variance $a_j\sigma$ ($j = 1, \ldots, m$). Under structures (a) and (b), the genetic effects exhibit at the burden score level for variants with MAFs$< 5\%$ and $< 1\%$, respectively, and the degree of (between-study) heterogeneity is measured by the coefficient of variation σ/θ. Under structure (c), the genetic effects exhibit at the individual variant level, and the degree of heterogeneity is measured by the ratio of variances σ/τ. For each percentage of potential causal variants and each genetic structure, we varied the degree of heterogeneity (i.e., σ/θ or σ/τ) from 0 to 2 with the increment of 0.5 and tuned the value of θ or τ such that the power is high enough to compare different methods.

Figures 4.1 and 4.2 display the power as a function of the degree of heterogeneity for the quantitative and binary traits, respectively. When the (between-study) heterogeneity is low, the FE tests (FE-BS, FE-VT, FE-VC and FE-VC-O) are more powerful than their RE counterparts (RE-BS, RE-VT, RE-VC and RE-VC-O), although the power loss of the latter is typically small. When the heterogeneity is high, the RE tests are much more powerful than the FE tests. Among the RE tests, RE-BS and RE-VT are the most powerful tests under structures (a) and (b), respectively, when the percentage of causal variants is high. Under structures (a) and (b) with low percentages of causal variants and under structure (c), RE-VC tends to be more powerful than RE-BS. The

Table 4.1: Type I error divided by the nominal significance level α for various meta-analysis methods

	Quantitative Phenotype			Binary Phenotype		
Tests	$\alpha = 10^{-2}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$	$\alpha = 10^{-2}$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$
FE-BS	0.99	0.98	0.96	1.00	0.98	0.93
RE-BS	0.99	0.95	1.02	0.99	0.93	0.92
FE-VT	0.99	0.97	1.05	0.99	0.94	0.93
RE-VT	0.99	0.97	0.94	0.96	0.87	0.81
FE-VC	1.00	0.96	0.97	0.98	0.95	1.00
RE-VC	0.98	0.96	0.95	0.93	0.91	0.95
FE-VC-O	0.99	1.03	0.92	1.00	0.97	0.96
RE-VC-O	1.00	1.04	1.04	0.96	0.99	0.90
Het-SKAT	0.95	0.86	0.82	0.85	0.77	0.69
Het-SKAT-O	1.00	0.89	0.78	0.96	0.87	0.61

power of RE-VC-O is near the top in all scenarios. Under structures (a) and (b) with low percentages of causal variants and under structure (c), RE-VC and RE-VC-O are considerably more powerful than Het-SKAT and Het-SKAT-O when the heterogeneity is low or moderate and have similar power to the latter when the heterogeneity is high.

We conducted another set of simulation studies by allowing genetic effects to exist in only a subset of the five studies. In such scenarios, it is sensible to test the association for each study and adjust the smallest p-value by the Bonferroni correction. Thus, we included this method, to be referred to as minP, in the simulation studies. We varied the the number of studies with genetic effects from 1 to 5 and set $\beta_k = \mu$ for those studies, where μ was generated under structure (a), (b) or (c). Figure 4.3 displays the results for the continuous trait when 50% of the variants are potentially causal. When the number of studies with genetic effects is 4 or 5, the RE tests are slightly less powerful than their FE counterparts. When the number is 1, 2 or 3, the RE tests are more powerful than the FE tests. The minP tests are less powerful than the RE tests except when the association exists in only one study. We also considered the binary trait and different percentages of causal variants, and the conclusions remain unchanged (data not shown).

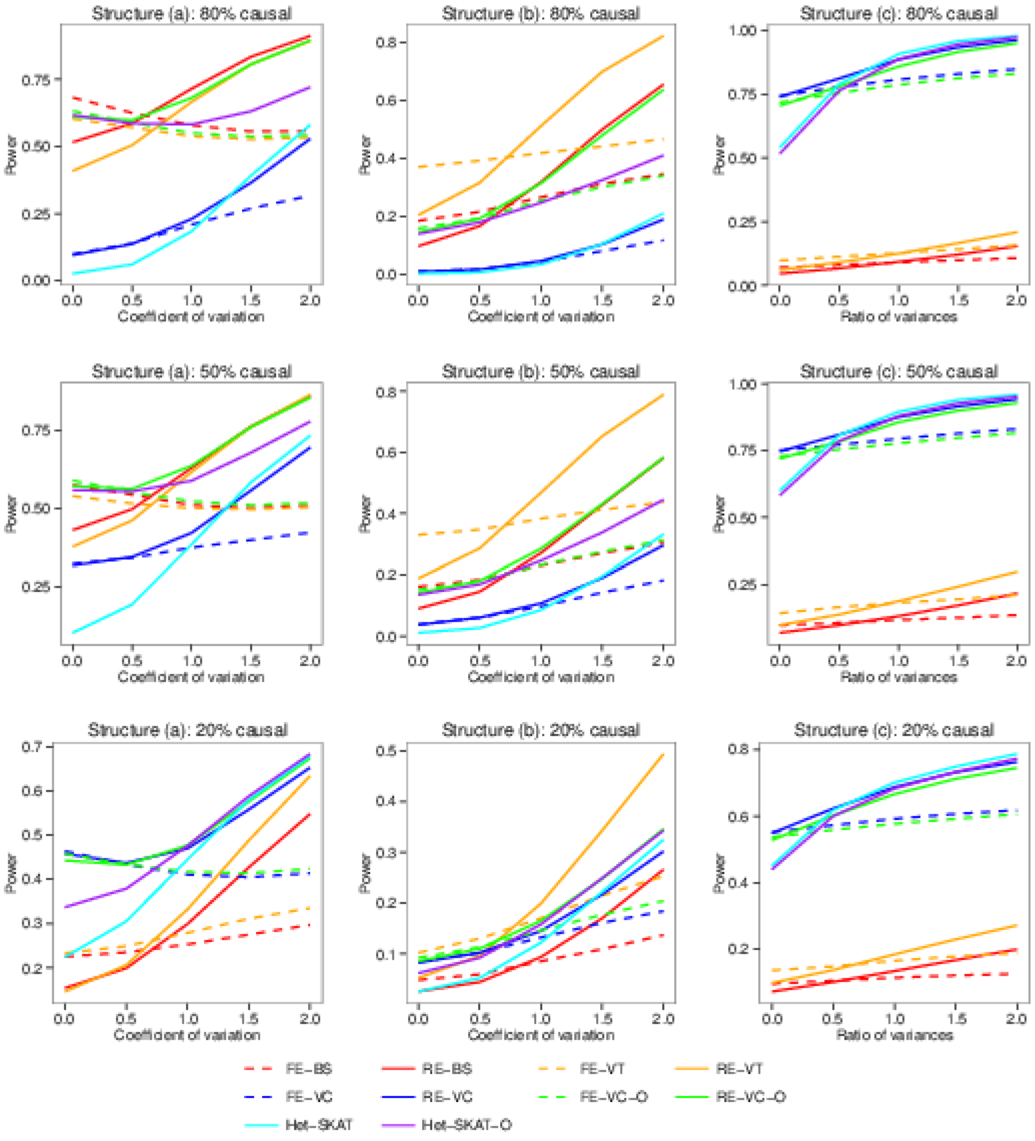


Figure 4.1: Power as a function of the between-study heterogeneity for the quantitative trait. The left, middle and right panels correspond to three different genetic structures: (a) genetic effects exhibit at the burden score level for variants with MAFs< 5%, (b) genetic effects exhibit at the burden score level for variants with MAFs< 1%, and (c) genetic effects exhibit at the individual variant level. For each structure, 80%, 50% or 20% of the variants in ten 300 base-pair regions were randomly selected to be potentially causal.

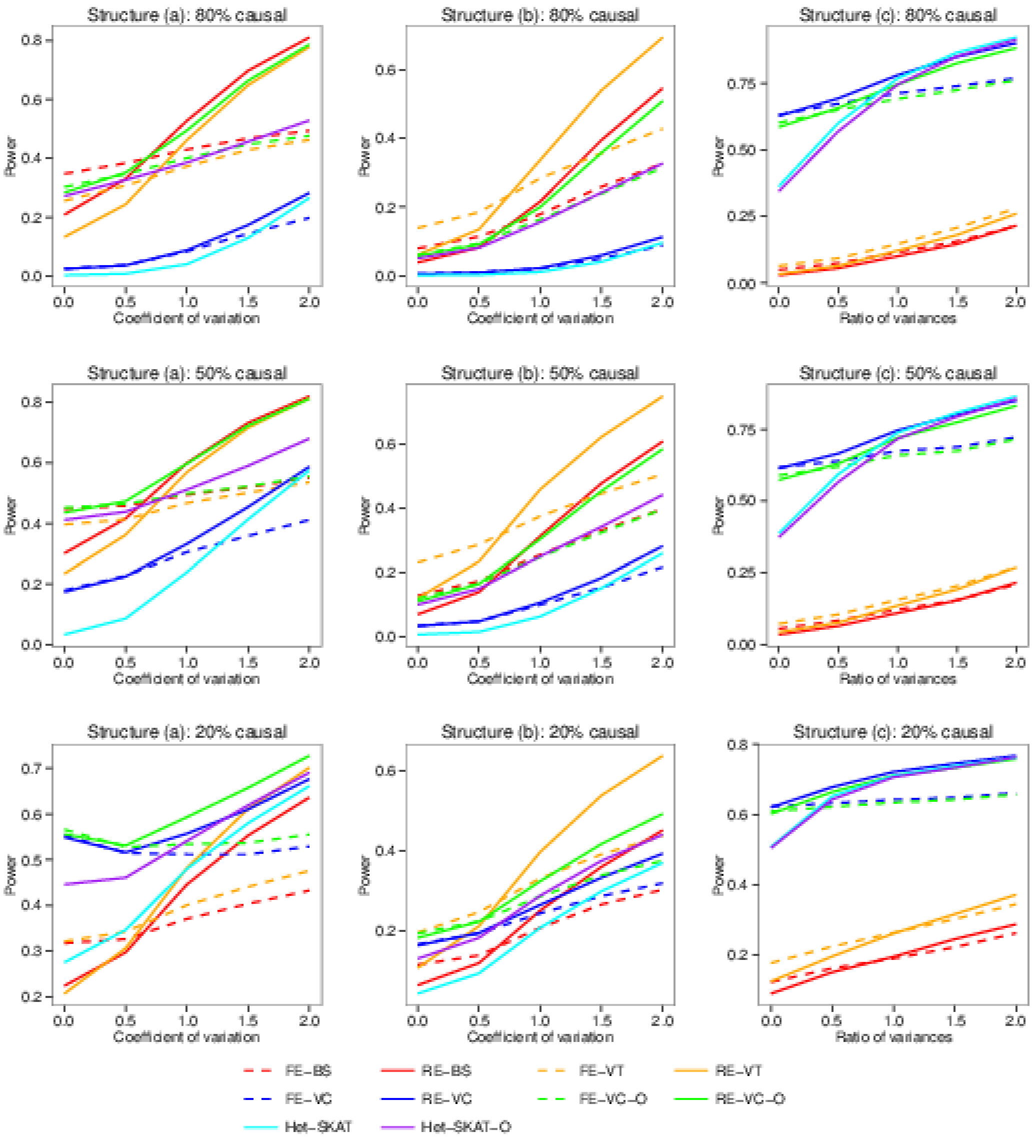


Figure 4.2: Power as a function of the between-study heterogeneity for the binary trait. The left, middle and right panels correspond to three different genetic structures: (a) genetic effects exhibit at the burden score level for variants with MAFs< 5%, (b) genetic effects exhibit at the burden score level for variants with MAFs< 1%, and (c) genetic effects exhibit at the individual variant level. For each structure, 80%, 50% or 20% of the variants in ten 300 base-pair regions were randomly selected to be potentially causal.

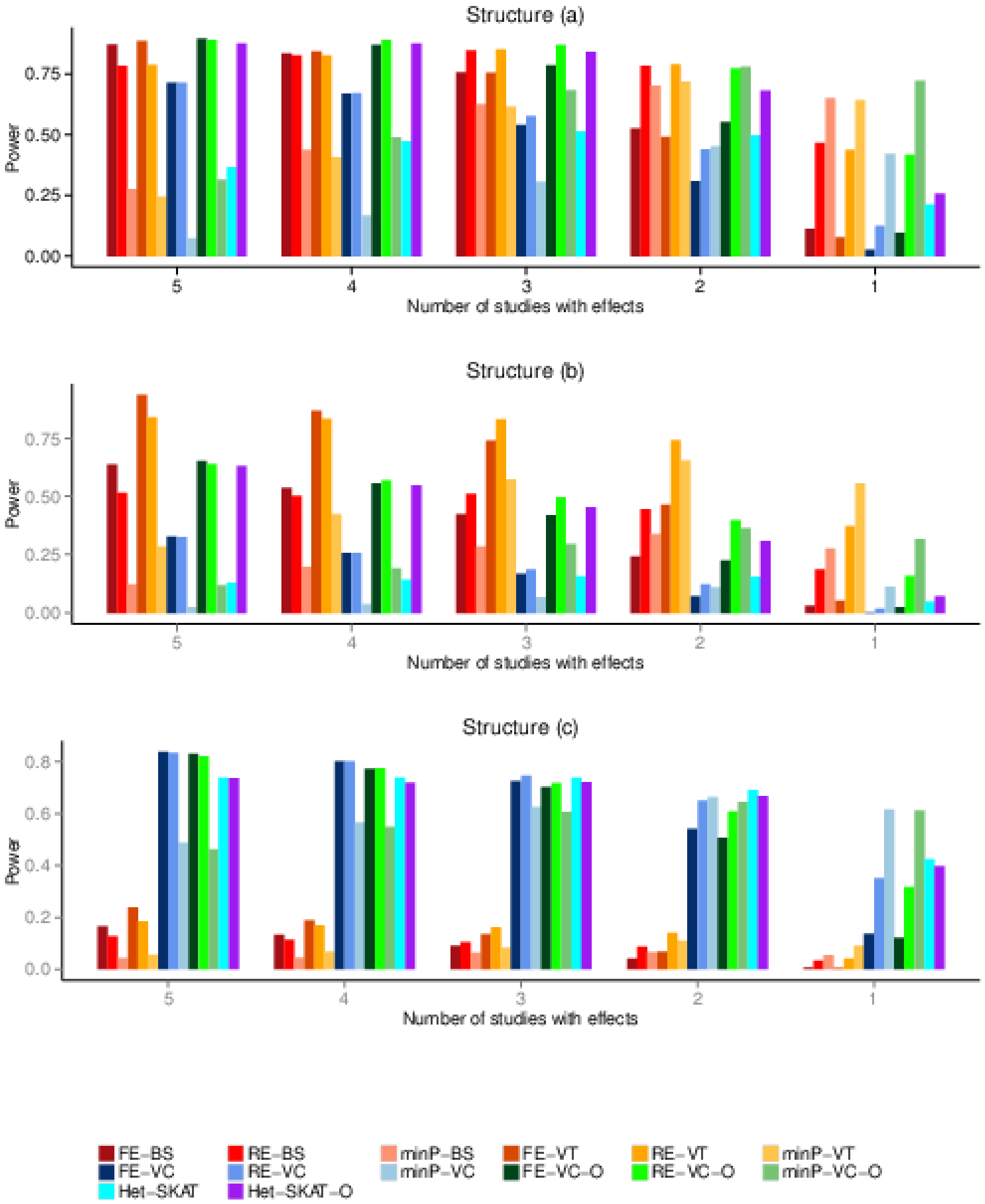


Figure 4.3: Power as a function of the number of studies with genetic effects on the quantitative trait. The upper, middle and lower panels correspond to three different genetic structures: (a) genetic effects exhibit at the burden score level for variants with MAFs$<$ 5%, (b) genetic effects exhibit at the burden score level for variants with MAFs$<$ 1%, and (c) genetic effects exhibit at the individual variant level. For each structure, 50% of the variants in ten 300 base-pair regions were randomly selected to be potentially causal.

4.4 Data analysis

The goal of the NHLBI ESP is to identify genetic variants in all protein-coding regions of the human genome that are associated with heart, lung, and blood diseases. The project consists of seven phenotype groups: low-density lipoprotein (LDL), body mass index (BMI), blood pressure (BP), early-onset myocardial infarction (EOMI), stroke, asthma and chronic obstructive pulmonary disease (COPD). In addition, there is a random sample of subjects who had measurements on a set of core variables (i.e., phenotypes, traits), which is referred to as deeply phenotyped reference (DPR). The DNA samples were sequenced on the Roche NimbleGen SeqCap EZ or Agilent SureSelect Human All Exon 50 MB at the University of Washington and the Broad Institute (Tennessen et al. 2012, Lang et al. 2014). The variants were called jointly at the University of Michigan. We set the individual genotype values to missing if the genotype depth was lower than 10. We restricted our attention to missense, nonsense and splice-site variants with call rates $> 90\%$ and MAFs$< 5\%$. We excluded any gene whose total minor allele count was less than 5 and ended up with a total of $14,878$ genes.

We considered LDL as the trait of interest and included several covariates in the linear regression: top two principal components for ancestry, age, age^2, gender, cohorts, and sequencing targets. The principal components were calculated from the sequencing data. The adjustment for sequencing targets was intended to remove potential batch effects. LDL was measured in the LDL, BMI, BP, EOMI, stroke and DPR groups, but not in the asthma and COPD groups. For each phenotype group, we treated the African American (AA) and European American (EA) samples separately. After excluding subjects with sex mismatch or relatedness, there were 296, 526, 214, 351, 75 and 240 AA subjects and 331, 0, 325, 484, 123 and 700 EA subjects in the LDL, BMI, BP, EOMI, stroke and DPR groups, respectively. In the meta-analysis, the score statistics for the eleven studies (i.e., phenotype group $\times$ race combinations) were obtained from SCORE-SeqTDS (Lin et al. 2013) and then combined to produce gene-level association tests. For the burden tests, we used the MAF thresholds of 1% and 5%, the corresponding tests being T1 and T5. The matrices B and W involved in the VC tests were specified in the same manner as in the simulation studies. We used 100,000 million Monte Carlo simulations to estimate the extreme p-values.

The results for T1, VT, VC and VC-O are displayed in Figure 4.4. (The results for T5 are similar to T1 and thus not shown. The burden scores for T1 and VT tests were unweighted; the weighted results are similar and not shown.) It is instructive to examine LDLR, which is the top gene in RE-T1. Several common variants in this gene were previously identified to be associated with lipid traits and coronary heart diseases, and heterogeneous associations among ethnic groups were reported (Zhang et al. 2013). In our data, there are 54 rare variants in LDLR, all with MAFs $< 1\%$, so the T1 and T5 tests are the same. In the T1 and VC-O tests, the RE meta-analysis provides stronger evidence of association than the FE meta-analysis: the RE-T1 and RE-VC-O p-values are 5.4×10^{-5} and 8.0×10^{-5}, respectively, whereas the FE-T1 and FE-VC-O p-values are 6.3×10^{-3} and 5.7×10^{-4}, respectively. The trend is reversed for the VT tests: the FE-VT and RE-VT p-values are 4.6×10^{-8} and 1.0×10^{-5}, respectively. For both FE-VT and RE-VT, the maxima of the test statistics occur at the MAF threshold of 0.02%. The forest plots shown in Figure 4.5 provide helpful insights. If we collapse variants with MAFs $< 0.02\%$, the effects of the burden scores are largely similar among the 11 studies; if we collapse variants with MAFs $< 1\%$, the effects of the burden scores are quite heterogeneous. As shown in Figure 4.6, for the variants with MAFs $< 0.02\%$, the carriers of mutations tend to have higher LDL levels than the non-carriers in all studies; for the variants with MAFs $> 0.02\%$, the distributions of the LDL values for the carriers are very different among studies. Figures 4.5 and 4.6 show that the heterogeneity is largely driven by the variability of genetic effects between AA and EA.

For LDLR, the RE-VC test (p-value $= 6.3 \times 10^{-4}$) is slightly less significant than the FE-VC test (p-value $= 3.1 \times 10^{-4}$). This is due to the fact that almost all the mutations of each variant are from one race group (see Figure 4.6), so the heterogeneity between the two races can be fully captured by the FE-VC test and the RE-VC test does not gain further information. By contrast, the RE-T1 test is more powerful than the FE-T1 test because there is considerable heterogeneity at the burden score level. The Het-SKAT and Het-SKAT-O p-values are 2.9×10^{-3} and 1.2×10^{-3}, which are much less significant than any of our RE tests. The minP p-values for T1, VT, VC and VC-O are 2.7×10^{-4}, 6.7×10^{-4}, 7.1×10^{-2} and 4.0×10^{-4}, respectively, which are less significant than their RE counterparts.

4.5 Discussion

In this article, we provide simple RE methods for all commonly used gene-level association tests, including the burden, VT and VC tests. Each test statistic contains contributions from both the mean and heterogeneity of the same type of genetic effect (i.e., at the burden score level for the burden and VT tests and at the individual variant level for the VC test). This is important because different tests are optimal for different scenarios. The RE tests are generally preferable to their FE counterparts because they are more powerful than the latter in the presence of moderate and high heterogeneity and have similar power to the latter when the heterogeneity is low, as demonstrated in the simulated and empirical data. The proposed methods are numerically stable and computationally efficient. They have been incorporated into the software MASS. It takes only a few minutes to conduct meta-analysis of several sequencing studies with thousands of genes.

For ethical and logistical reasons, summary statistics are more readily available than individual participant data. The proposed methods are based on score statistics and are

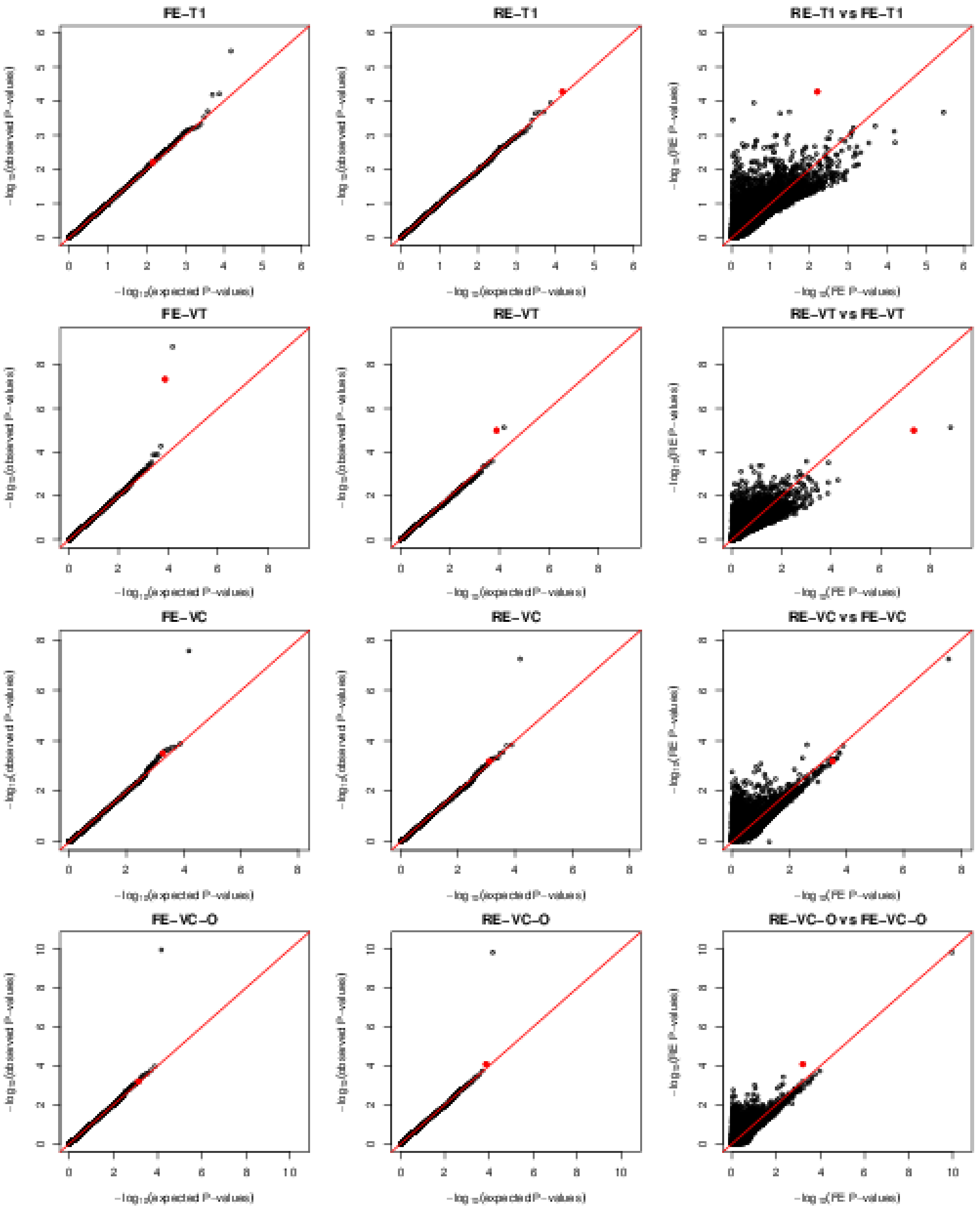
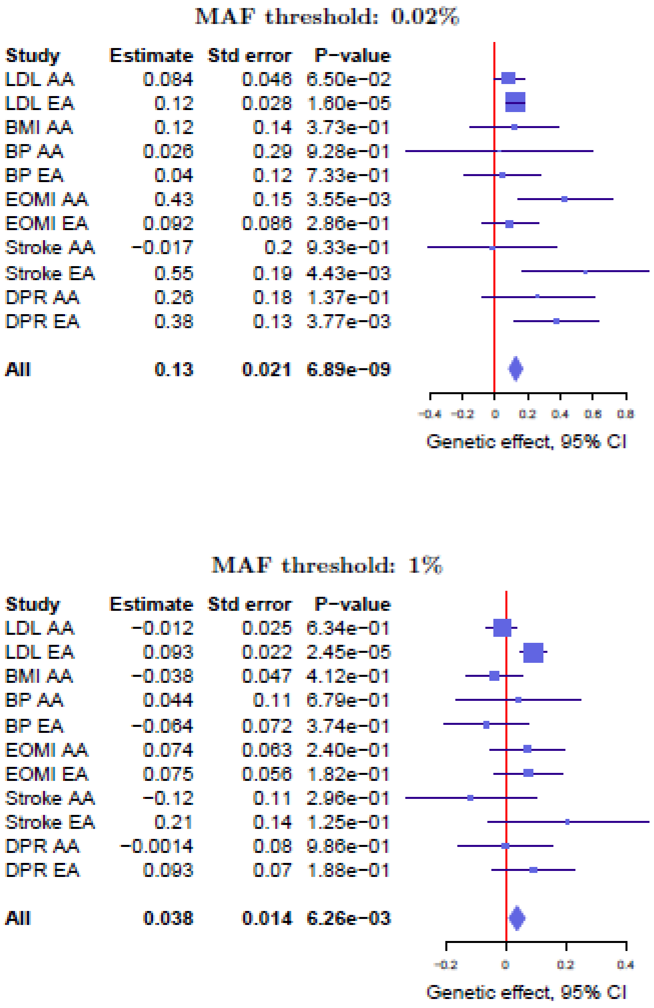


Figure 4.4: Meta-analysis of the eleven studies in the NHLBI ESP: the left and middle panels are the quantile-quantile plots for the FE and RE tests, and the right panel compares the RE and FE results. The red dot indicates the gene LDLR.



MAF threshold: 0.02%

Study	Estimate	Std error	P−value
LDL AA	0.084	0.046	6.50e−02
LDL EA	0.12	0.028	1.60e−05
BMI AA	0.12	0.14	3.73e−01
BP AA	0.026	0.29	9.28e−01
BP EA	0.04	0.12	7.33e−01
EOMI AA	0.43	0.15	3.55e−03
EOMI EA	0.092	0.086	2.86e−01
Stroke AA	−0.017	0.2	9.33e−01
Stroke EA	0.55	0.19	4.43e−03
DPR AA	0.26	0.18	1.37e−01
DPR EA	0.38	0.13	3.77e−03
All	**0.13**	**0.021**	**6.89e−09**

MAF threshold: 1%

Study	Estimate	Std error	P−value
LDL AA	−0.012	0.025	6.34e−01
LDL EA	0.093	0.022	2.45e−05
BMI AA	−0.038	0.047	4.12e−01
BP AA	0.044	0.11	6.79e−01
BP EA	−0.064	0.072	3.74e−01
EOMI AA	0.074	0.063	2.40e−01
EOMI EA	0.075	0.056	1.82e−01
Stroke AA	−0.12	0.11	2.96e−01
Stroke EA	0.21	0.14	1.25e−01
DPR AA	−0.0014	0.08	9.86e−01
DPR EA	0.093	0.07	1.88e−01
All	**0.038**	**0.014**	**6.26e−03**

Figure 4.5: Forest plots for the burden tests with two MAF thresholds for the gene LDLR in the NHLBI ESP. For each study, the estimate of the genetic effect is shown by the square and the corresponding 95% confidence interval is shown by the line. The meta-estimate of the genetic effect and the corresponding 95% confidence interval are shown by the diamond.

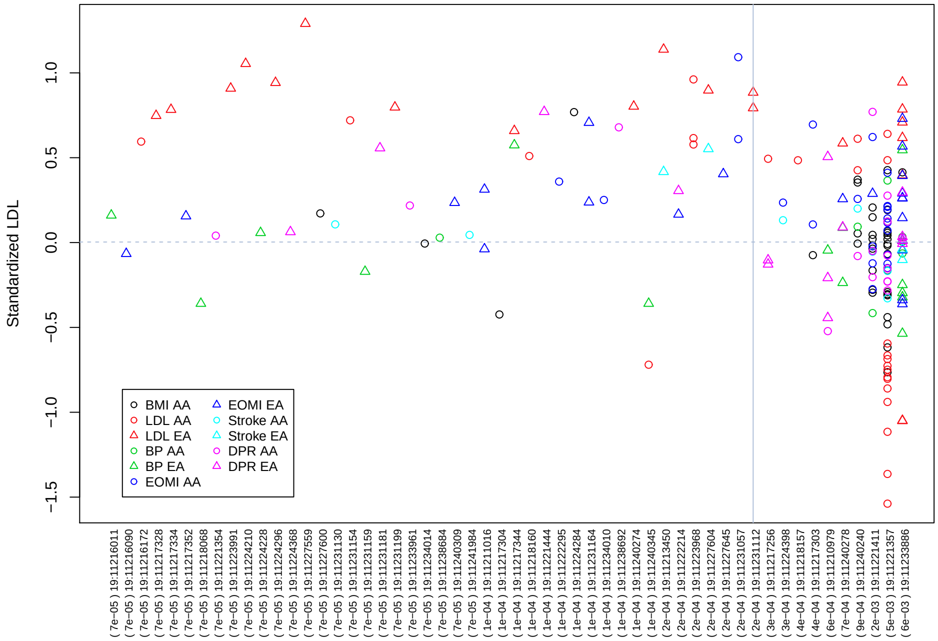


Figure 4.6: Standardized LDL values for the carriers of the LDLR mutations in the NHLBI ESP. Each point represents an individual who carries a mutation. There are 54 polymorphic sites with the chr:position IDs and MAFs labeled on the x-axis. The variants are ordered by the MAFs. The vertical line pertains to the MAF threshold at which the test statistics of FE-VT and RE-VT are maximized. The five phenotype groups are indicated by different colors. AA and EA subjects are shown in circles and triangles, respectively. The horizontal line pertains to the average LDL value among individuals who do not carry any mutation.

as efficient as joint analysis of individual participant data (Lin and Zeng 2010). Since it inputs score statistics rather than individual participant data, our framework can accommodate any phenotype and any study design or studies with different designs. For sequencing studies, score statistics are preferable to Wald and likelihood ratio statistics (Lin and Tang 2011). The latter would entail estimation of the between-study variance (in the univariate case) or covariance matrix (in the multivariate case), which is numerically unstable or infeasible for rare variants.

The conventional RE approach is focused on the mean effect size (DerSimonian and Laird 1986). That approach is not suitable for association testing for several reasons. First, it tests the null hypothesis of no mean effect while allowing between-study heterogeneity. This is not the relevant hypothesis for association testing because the existence of heterogeneity implies association in at least some studies. For this reason, conventional RE tests are almost always less significant than FE tests and thus have rarely been used in genetic association studies (Han and Eskin 2011). Second, the conventional RE approach is based on the asymptotic distribution, which requires a large number of studies, but the number of sequencing studies is usually small. Third, existing multivariate RE methods leave the between-study covariance matrix completely un-structured (Jackson et al. 2010, Chen et al. 2012) and thus may lose power due to the large number of degrees of freedom.

Our approach reflects the spirit of Han and Eskin (2011) in that it tests the joint null hypothesis that there is no mean effect and no between-study heterogeneity. Our framework differs from Han and Eskin's in three major aspects. First, their method is restricted to single-variant analysis of common SNPs whereas our methods deal with gene-level tests of rare variants. Second, their test statistic is univariate whereas our framework accommodates both univariate and multivariate test statistics. Third, their method is based on the likelihood ratio statistic whereas our methods are based on the score statistic.

Our RE tests were derived under random-effects models and may appear to rely on the normality of random effects, which is an untestable assumption. However, the random-effects models were only used to motivate the forms of the test statistics. By using Monte Carlo simulation rather than asymptotic approximation to obtain the p-values, the proposed tests have correct type I error even if the underlying random-effects models are completely wrong.

Our framework can be readily extended to handle multiple levels of between-study heterogeneity. Suppose that there are several (ancestry) groups of studies such that the genetic effects are homogeneous within each group but heterogeneous across groups. In that case, we will sum the score statistics and information matrices over the studies within each group and then construct the RE test statistics to account for the between-group heterogeneity.

If the burden score is created under the additive mode of inheritance as the sum or a weighted sum of the mutation counts over the variant sites (Madsen and Browning 2009, Morris and Zeggini 2010, Price et al. 2010, Lin and Tang 2011), then the score statistics and information matrices for the burden and VT tests can be generated from the score vector and information matrix for testing individual invariants used in the VC tests (Hu et al. 2013). Specifically, the score statistic for the burden test is the sum or a weighted sum of the score statistics for testing the effects of individual variants. Under the dominant mode of inheritance, the burden score indicates whether there is any mutation among the variant sites (Morgenthaler and Thilly 2007, Li and Leal 2008). Then the above conversion can no longer be used. Our framework allows any mode of inheritance since the creation of the burden scores is external to the construction of the test statistics.

In meta-analysis, it is wise to have consistency across studies in terms of quality control criteria, gene annotation, variant selection and MAF estimation. This requirement is less essential for the RE tests than for the FE tests because heterogeneity (of genetic effects) is allowed for the former but not for the latter. For studies that use different exome capturing kits or studies in which some use whole-exome sequencing while others use exome chips, the variants captured can be quite different among studies. In such situations, the RE tests should be used since the effects are expected to be heterogeneous.

APPENDIX : Test statistics in Chapter 4

Let $\widehat{\beta}_k$ denote the maximum likelihood estimator (MLE) of β_k. By the MLE theory, $\widehat{\beta}_k$ is approximately normal with mean β_k and covariance matrix $\mathcal{I}_k^{-1}$, where $\mathcal{I}_k$ is the (profile) information matrix for β_k. Under model4.1 with fixed μ, $\widehat{\beta}_k$ is approximately normal with mean μ and covariance matrix $\mathcal{I}_k^{-1} + \Sigma$. For rare variants, the effect estimators $\widehat{\beta}_k$'s are unstable and may not be computable. Thus, we construct test statistics based on the score statistics rather than the Wald or likelihood ratio statistics. We use the scaled score statistic $X_k = V_k^{-1}U_k$ as a surrogate for $\widehat{\beta}_k$. Note that V_k is the same as $\mathcal{I}_k$ except that the former is the (profile) information matrix evaluated at $\beta_k = 0$ and the latter at $\widehat{\beta}_k$. For small β_k, the statistic X_k is approximately normal with mean μ and covariance matrix $\Omega_k = V_k^{-1} + \sigma B$,

The log-likelihood function for μ and σ based on the statistics X_k $(k = 1, \cdots, K)$ can be written as

$$l(\mu, \sigma) = -\frac{1}{2}\sum_{k=1}^{K}(X_k - \mu)^{\mathrm{T}}\Omega_k^{-1}(X_k - \mu) - \frac{1}{2}\sum_{k=1}^{K}\log|\Omega_k|.$$

By tedious but straightforward matrix differentiation, we can show that the score function consists of

$$S_\mu(\mu, \sigma) = \frac{\partial l(\mu, \sigma)}{\partial \mu} = \sum_{k=1}^{K}\Omega_k^{-1}(X_k - \mu),$$

$$S_\sigma(\mu, \sigma) = \frac{\partial l(\mu, \sigma)}{\partial \sigma} = \frac{1}{2}\sum_{k=1}^{K}(X_k - \mu)^{\mathrm{T}}\Omega_k^{-1}B\Omega_k^{-1}(X_k - \mu) - \frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}\left(\Omega_k^{-1}B\right),$$

and the corresponding Fisher information matrix is

$$\mathcal{I}(\mu, \sigma) = -E\begin{bmatrix}\frac{\partial^2 l(\mu,\sigma)}{\partial\mu\partial\mu^{\mathrm{T}}} & \frac{\partial^2 l(\mu,\sigma)}{\partial\mu\partial\sigma} \\ \frac{\partial^2 l(\mu,\sigma)}{\partial\sigma\partial\mu^{\mathrm{T}}} & \frac{\partial^2 l(\mu,\sigma)}{\partial\sigma\partial\sigma}\end{bmatrix} = \begin{bmatrix}\sum_{k=1}^{K}\Omega_k^{-1} & 0 \\ 0 & \frac{1}{2}\mathrm{tr}\left(\sum_{k=1}^{K}\Omega_k^{-1}B\Omega_k^{-1}B\right)\end{bmatrix}.$$

The score statistic for testing $H_0 : \mu = 0$ and $\sigma = 0$ is

$$\begin{bmatrix} U_\mu^{\mathrm{T}} & U_\sigma \end{bmatrix} V_{\mu\sigma}^{-1} \begin{bmatrix} U_\mu \\ U_\sigma \end{bmatrix} = U_\mu^{\mathrm{T}} V_\mu^{-1} U_\mu + \frac{U_\sigma^2}{V_\sigma},$$

where $U_\mu = S_\mu(0,0) = \sum_{k=1}^K U_k$, $U_\sigma = S_\sigma(0,0) = \frac{1}{2}\sum_{k=1}^K U_k^{\mathrm{T}} B U_k - \frac{1}{2}\sum_{k=1}^K \mathrm{tr}(V_k B)$, $V_{\mu\sigma} = \mathcal{I}(0,0)$, $V_\mu = \sum_{k=1}^K V_k$, and $V_\sigma = \frac{1}{2}\mathrm{tr}\big(\sum_{k=1}^K V_k B V_k B\big)$.

We now assume that μ is normal with mean 0 and covariance matrix τW. Write $X = (X_1^{\mathrm{T}}, \cdots, X_K^{\mathrm{T}})^{\mathrm{T}}$. The statistic X is approximately normal with mean 0 and covariance matrix $\Omega = \tau(J_K \otimes W) + \sigma(I_K \otimes B) + \mathrm{diag}(V_1^{-1}, \cdots, V_K^{-1})$, where J_K is a $K \times K$ matrix composed of 1, I_K is a K-dimensional identity matrix, and $\otimes$ is the Kronecker product. Then the log-likelihood function for τ and σ can be written as

$$l(\tau, \sigma) = -\frac{1}{2} X^{\mathrm{T}} \Omega^{-1} X - \frac{1}{2} \log |\Omega|.$$

The score function consists of

$$S_\tau(\tau, \sigma) = \frac{\partial l(\tau,\sigma)}{\partial \tau} = \frac{1}{2} X^{\mathrm{T}} \Omega^{-1} (J_K \otimes W) \Omega^{-1} X - \frac{1}{2}\mathrm{tr}\big(\Omega^{-1}(J_K \otimes W)\big),$$

$$S_\sigma(\tau, \sigma) = \frac{\partial l(\tau,\sigma)}{\partial \sigma} = \frac{1}{2} X^{\mathrm{T}} \Omega^{-1} (I_K \otimes B) \Omega^{-1} X - \frac{1}{2}\mathrm{tr}\big(\Omega^{-1}(I_K \otimes B)\big),$$

and the corresponding Fisher information matrix is

$$\mathcal{I}(\tau, \sigma) = -E \begin{bmatrix} \frac{\partial^2 l(\tau,\sigma)}{\partial\tau\partial\tau} & \frac{\partial^2 l(\tau,\sigma)}{\partial\tau\partial\sigma} \\ \frac{\partial^2 l(\tau,\sigma)}{\partial\sigma\partial\tau} & \frac{\partial^2 l(\tau,\sigma)}{\partial\sigma\partial\sigma} \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} \mathrm{tr}\big(\Omega^{-1}(J_K \otimes W)\Omega^{-1}(J_K \otimes W)\big) & \mathrm{tr}\big(\Omega^{-1}(J_K \otimes W)\Omega^{-1}(I_K \otimes B)\big) \\ \mathrm{tr}\big(\Omega^{-1}(I_K \otimes B)\Omega^{-1}(J_K \otimes W)\big) & \mathrm{tr}\big(\Omega^{-1}(I_K \otimes B)\Omega^{-1}(I_K \otimes B)\big) \end{bmatrix}.$$

The score statistic for testing $H_0 : \tau = \sigma = 0$ is

$$\begin{bmatrix} U_\tau & U_\sigma \end{bmatrix} V_{\tau\sigma}^{-1} \begin{bmatrix} U_\tau \\ U_\sigma \end{bmatrix},$$

where $U_\tau = S_\tau(0,0) = \frac{1}{2}U_\mu^{\mathrm{T}} W U_\mu - \frac{1}{2}\mathrm{tr}(V_\mu W)$, $U_\sigma = S_\sigma(0,0) = \frac{1}{2}\sum_{k=1}^{K} U_k^{\mathrm{T}} B U_k - \frac{1}{2}\mathrm{tr}(V_\mu B)$, and

$$V_{\tau\sigma} = \mathcal{I}(0,0) = \frac{1}{2}\begin{bmatrix} \mathrm{tr}\big(V_\mu W V_\mu W\big) & \mathrm{tr}\big(\sum_{k=1}^{K} V_k W V_k B\big) \\ \mathrm{tr}\big(\sum_{k=1}^{K} V_k B V_k W\big) & \mathrm{tr}\big(\sum_{k=1}^{K} V_k B V_k B\big) \end{bmatrix}.$$

REFERENCE

Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S., Yosef, N., Ruppin, E., Sharan, R., Vaisse, C., Sunyaev, S., Dent, R., Cohen, J., McPherson, R., and Pennacchio, L. a. (2007), "Medical sequencing at the extremes of human body mass.," *Am. J. Hum. Genet.*, 80(4), 779–91.

Chen, H., Manning, A. K., and Dupuis, J. (2012), "A method of moments estimator for random effect multivariate meta-analysis.," *Biometrics*, 68(4), 1278–84.

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(5685), 869–72.

DerSimonian, R., and Laird, N. (1986), "Meta-analysis in clinical trials.," *Contr. Clin. Trials.*, 7(3), 177–188.

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. a., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., Latiano, A., Mathew, C. G., Montgomery, G. W., Prescott, N. J., Raychaudhuri, S., Rotter, J. I., Schumm, P., Sharma, Y., Simms, L. a., Taylor, K. D., Whiteman, D., Wijmenga, C., Baldassano, R. N., Barclay, M., Bayless, T. M., Brand, S., Büning, C., Cohen, A., Colombel, J.-F., Cottone, M., Stronati, L., Denson, T., De Vos, M., D'Inca, R., Dubinsky, M., Edwards, C., Florin, T., Franchimont, D., Gearry, R., Glas, J., Van Gossum, A., Guthery, S. L., Halfvarson, J., Verspaget, H. W., Hugot, J.-P., Karban, A., Laukens, D., Lawrance, I., Lemann, M., Levine, A., Libioulle, C., Louis, E., Mowat, C., Newman, W., Panés, J., Phillips, A., Proctor, D. D., Regueiro, M., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Seibold, F., Steinhart, a. H., Stokkers, P. C. F., Torkvist, L., Kullak-Ublick, G., Wilson, D., Walters, T., Targan, S. R., Brant, S. R., Rioux, J. D., D'Amato, M., Weersma, R. K., Kugathasan, S., Griffiths, A. M., Mansfield, J. C., Vermeire, S., Duerr, R. H., Silverberg, M. S., Satsangi, J., Schreiber, S., Cho, J. H., Annese, V., Hakonarson, H., Daly, M. J., and Parkes, M. (2010), "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.," *Nat. Genet.*, 42(12), 1118–25.

Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R., and Amos, C. I. (2008), "Shifting Paradigm of Association Studies : Value of Rare Single-Nucleotide Polymorphisms," *Am. J. Hum. Genet.*, (January), 100–112.

Han, B., and Eskin, E. (2011), "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies.," *Am. J. Hum. Genet.*, 88(5), 586–598.

Han, F., and Pan, W. (2010), "A data-adaptive sum test for disease association with multiple common or rare variants.," *Human heredity*, 70(1), 42–54.

Hardy, J., and Singleton, A. (2009), "Genomewide association studies and human disease.," *N. Engl. J. Med.*, 360(17), 1759–68.

Heid, I. M., Huth, C., Loos, R. J. F., Kronenberg, F., Adamkova, V., Anand, S. S., Ardlie, K., Biebermann, H., Bjerregaard, P., Boeing, H., Bouchard, C., Ciullo, M., Cooper, J. a., Corella, D., Dina, C., Engert, J. C., Fisher, E., Francès, F., Froguel, P., Hebebrand, J., Hegele, R. a., Hinney, A., Hoehe, M. R., Hu, F. B., Hubacek, J. a., Humphries, S. E., Hunt, S. C., Illig, T., Järvelin, M.-R., Kaakinen, M., Kollerits, B., Krude, H., Kumar, J., Lange, L. a., Langer, B., Li, S., Luchner, A., Lyon, H. N., Meyre, D., Mohlke, K. L., Mooser, V., Nebel, A., Nguyen, T. T., Paulweber, B., Perusse, L., Qi, L., Rankinen, T., Rosskopf, D., Schreiber, S., Sengupta, S., Sorice, R., Suk, A., Thorleifsson, G., Thorsteinsdottir, U., Völzke, H., Vimaleswaran, K. S., Wareham, N. J., Waterworth, D., Yusuf, S., Lindgren, C., McCarthy, M. I., Lange, C., Hirschhorn, J. N., Laird, N., and Wichmann, H.-E. (2009), "Meta-analysis of the INSIG2 association with obesity including 74,345 individuals: does heterogeneity of estimates relate to study design?," *PLoS Genet.*, 5(10), e1000694.

Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M. C., Speliotes, E. K., Mägi, R., Workalemahu, T., White, C. C., Bouatia-Naji, N., Harris, T. B., Berndt, S. I., Ingelsson, E., Willer, C. J., Weedon, M. N., Luan, J., Vedantam, S., Esko, T. o., Kilpeläinen, T. O., Kutalik, Z., Li, S., Monda, K. L., Dixon, A. L., Holmes, C. C., Kaplan, L. M., Liang, L., Min, J. L., Moffatt, M. F., Molony, C., Nicholson, G., Schadt, E. E., Zondervan, K. T., Feitosa, M. F., Ferreira, T., Lango Allen, H., Weyant, R. J., Wheeler, E., Wood, A. R., Estrada, K., Goddard, M. E., Lettre, G., Mangino, M., Nyholt, D. R., Purcell, S., Smith, A. V., Visscher, P. M., Yang, J., McCarroll, S. a., Nemesh, J., Voight, B. F., Absher, D., Amin, N., Aspelund, T., Coin, L., Glazer, N. L., Hayward, C., Heard-Costa, N. L., Hottenga, J.-J., Johansson, A., Johnson, T., Kaakinen, M., Kapur, K., Ketkar, S., Knowles, J. W., Kraft, P., Kraja, A. T., Lamina, C., Leitzmann, M. F., McKnight, B., Morris, A. P., Ong, K. K., Perry, J. R. B., Peters, M. J., Polasek, O., Prokopenko, I., Rayner, N. W., Ripatti, S., Rivadeneira, F., Robertson, N. R., Sanna, S., Sovio, U., Surakka, I., Teumer, A., van Wingerden, S., Vitart, V., Zhao, J. H., Cavalcanti-Proença, C., Chines, P. S., Fisher, E., Kulzer, J. R., Lecoeur, C., Narisu, N., Sandholt, C., Scott, L. J., Silander, K., Stark, K., Tammesoo, M.-L., Teslovich, T. M., Timpson, N. J., Watanabe, R. M., Welch, R., Chasman, D. I., Cooper, M. N., Jansson, J.-O., Kettunen, J., Lawrence, R. W., Pellikka, N., Perola, M., Vandenput, L., Alavere, H., Almgren, P., Atwood, L. D., Bennett, A. J., Biffar, R., Bonnycastle, L. L., Bornstein, S. R., Buchanan, T. a., Campbell, H., Day, I. N. M., Dei, M., Dörr, M., Elliott, P., Erdos, M. R., Eriksson, J. G., Freimer, N. B., Fu, M., Gaget, S., Geus, E. J. C., Gjesing, A. P., Grallert, H., Grässler, J., Groves, C. J., Guiducci, C., Hartikainen, A.-L., Hassanali, N., Havulinna, A. S., Herzig, K.-H., Hicks, A. a., Hui, J., Igl, W., Jousilahti, P., Jula, A., Kajantie, E., Kinnunen, L., Kolcic, I., Koskinen, S., Kovacs, P., Kroemer, H. K., Krzelj, V., Kuusisto, J., Kvaloy, K., Laitinen, J., Lantieri, O., Lathrop, G. M., Lokki, M.-L., Luben, R. N., Ludwig, B., McArdle, W. L., McCarthy, A., Morken, M. a., Nelis, M., Neville, M. J., Paré, G., Parker, A. N., Peden, J. F., Pichler, I., Pietiläinen, K. H., Platou, C. G. P., Pouta, A., Ridderstrå le, M., Samani, N. J., Saramies, J., Sinisalo, J., Smit, J. H., Strawbridge, R. J., Stringham, H. M., Swift, A. J., Teder-Laving, M., Thomson, B., Usala, G., van Meurs, J. B. J., van Ommen, G.-J., Vatin, V., Volpato, C. B., Wallaschofski, H., Walters, G. B., Widen, E., Wild, S. H., Willemsen, G., Witte, D. R., Zgaga, L., Zitting, P., Beilby, J. P., James, A. L., Kähönen, M., Lehtimäki, T., Nieminen, M. S., Ohlsson, C., Palmer, L. J., Raitakari, O., Ridker, P. M., Stumvoll, M., Tönjes, A., Viikari, J., Balkau, B., Ben-Shlomo, Y., Bergman, R. N., Boeing, H., Smith, G. D., Ebrahim, S., Froguel, P., Hansen, T., Hengstenberg, C., Hveem, K., Isomaa, B., Jø rgensen, T., Karpe, F., Khaw, K.-T., Laakso, M., Lawlor, D. a., Marre, M., Meitinger, T., Metspalu, A., Midthjell, K., Pedersen, O., Salomaa, V., Schwarz, P. E. H., Tuomi, T., Tuomilehto, J., Valle, T. T., Wareham, N. J., Arnold, A. M., Beckmann, J. S., Bergmann, S., Boerwinkle, E., Boomsma, D. I., Caulfield, M. J., Collins, F. S., Eiriksdottir, G., Gudnason, V., Gyllensten, U., Hamsten, A., Hattersley, A. T., Hofman, A., Hu, F. B., Illig, T., Iribarren, C., Jarvelin, M.-R., Kao, W. H. L., Kaprio, J., Launer, L. J., Munroe, P. B., Oostra, B., Penninx, B. W., Pramstaller, P. P., Psaty, B. M., Quertermous, T., Rissanen, A., Rudan, I., Shuldiner, A. R., Soranzo, N., Spector, T. D., Syvanen, A.-C., Uda, M., Uitterlinden, A., Völzke, H., Vollenweider, P., Wilson, J. F., Witteman, J. C., Wright, A. F., Abecasis, G. R., Boehnke, M., Borecki, I. B., Deloukas, P., Frayling, T. M., Groop, L. C., Haritunians, T., Hunter, D. J., Kaplan, R. C., North, K. E., O'Connell, J. R., Peltonen, L., Schlessinger, D., Strachan, D. P., Hirschhorn, J. N., Assimes, T. L., Wichmann, H.-E., Thorsteinsdottir, U., van Duijn, C. M., Stefansson, K., Cupples, L. A., Loos, R. J. F., Barroso, I., McCarthy, M. I., Fox, C. S., Mohlke, K. L., and Lindgren, C. M. (2010), "Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution.," *Nature Genet.*, 42(11), 949–60.

Hu, Y.-J., Berndt, S. I., Gustafsson, S., Ganna, A., Hirschhorn, J., North, K. E., Ingelsson, E., and Lin, D.-Y. (2013), "Meta-analysis of gene-level associations for rare variants based on single-variant statistics.," *Am. J. Hum. Genet.*, 93, 236–248.

Hudson, R. R. (2002), "Generating samples under a Wright-Fisher neutral model of genetic variation," *Bioinformatics*, 18(2), 337–338.

Ioannidis, J. P., Patsopoulos, N. a., and Evangelou, E. (2007), "Heterogeneity in Meta-Analyses of Genome-Wide Association Investigations," *PLoS ONE*, 2(9), e841.

Jackson, D., White, I. R., and Thompson, S. G. (2010), "Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses.," *Statistics in medicine*, 29(12), 1282–97.

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.," *Proc Natl Acad Sci USA*, 106(10), 3871–3876.

Lang, L. A., Hu, Y., Zhang, H., Xue, C., Schmidt, E. M., Tang, Z., Bizon, C., Lange, E. M., Smith, J. D., Turner, E. H., and et al. (2014), "Novel rare and low frequency coding variants associated with LDL cholesterol.," *Am J Hum Genet*, in press.

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

Lee, S., Teslovich, T., Boehnke, M., and Lin, X. (2013), "General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies," *Am. J. Hum. Genet.*, pp. 1–12.

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

Lin, D.-Y. (2005), "An efficient Monte Carlo approach to assessing statistical significance in genomic studies.," *Bioinformatics*, 21(6), 781–787.

Lin, D.-Y., and Tang, Z.-Z. (2011), "A general framework for detecting disease associations with rare variants in sequencing studies.," *Am. J. Hum. Genet.*, 89(3), 354–67.

Lin, D.-Y., and Zeng, D. (2006), "Likelihood-based inference on haplotype effects in genetic association studies (with discussion)," *J. Am. Stat. Ass.*, 101, 89118.

Lin, D.-Y., and Zeng, D. (2010), "On the relative efficiency of using summary statistics versus individual-level data in meta-analysis.," *Biometrika*, 97(2), 321–332.

Lin, D.-Y., Zeng, D., and Tang, Z.-Z. (2013), "Quantitative trait analysis in sequencing studies under trait-dependent sampling.," *Proc. Natl. Acad. Sci. USA*, pp. 1–6.

Liu, D. J., and Leal, S. M. (2010), "A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions.," *PLoS genetics*, 6(10), e1001156.

Liu, D. J., Peloso, G. M., Zhan, X., Oddgeir, H., Zawistowski, M., Feng, S., Nikpay, M., Auer, P. L., Goel, A., Zhang, H., Peters, U., Farrall, M., Orho-Melander, M., Kooperberg, C., McPherson, R., Watkins, H., Willer, C. J., Hveem, K., Melander, O., Kathiresan, S., and Abecasis, G. R. (2014), "Meta-analysis of gene-level tests for rare variant association.," *Nat Genet*, 46, 200–204.

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

Morgenthaler, S., and Thilly, W. G. (2007), "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST).," *Mutat. Res.*, 615(1-2), 28–56.

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

Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J. A. (2009), "Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes.," *Science*, 333(April), 387–389.

Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L.-J., and Sunyaev, S. R. (2010), "Pooled association tests for rare variants in exon-resequencing studies.," *Am. J. Hum. Genet.*, 86(6), 832–8.

Pritchard, J. K. (2001), "Are Rare Variants Responsible for Susceptibility to Complex Diseases ?," *Am. J. Hum. Genet.*, (1), 124–137.

Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. (2012), "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.," *Nucleic acids research*, 40, D130–5.

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I. W., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., Hughes, T. E., Groop, L., Altshuler, D., Almgren, P., Florez, J. C., Meyer, J., Ardlie, K., Bengtsson Boström, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H. N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Rå stam, L., Speliotes, E. K., Taskinen, M.-R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjögren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S. B., Chirn, G.-W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., and Purcell, S. (2007), "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.," *Science*, 316(5829), 1331–6.

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

Scott, L. J., Muglia, P., Kong, X. Q., Guan, W., Flickinger, M., Upmanyu, R., Tozzi, F., Li, J. Z., Burmeister, M., Absher, D., Thompson, R. C., Francks, C., Meng, F., Antoniades, A., Southwick, A. M., Schatzberg, A. F., Bunney, W. E., Barchas, J. D., Jones, E. G., Day, R., Matthews, K., Mcguffin, P., Strauss, J. S., Kennedy, J. L., Middleton, L., Roses, A. D., Watson, S. J., Vincent, J. B., Myers, R. M., Farmer, A. E., Akil, H., Burns, D. K., and Boehnke, M. (2009), "Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry," *Proc. Natl. Acad. Sci. USA*, 106(18).

Sun, J., Zheng, Y., and Hsu, L. (2013), "A unified mixed-effects model for rare-variant association in sequencing studies.," *Genetic epidemiology*, 37(4), 334–44.

Tang, Z.-Z., and Lin, D.-Y. (2013), "MASS : meta-analysis of score statistics for sequencing studies," *Bioinformatics*, pp. 1–3.

Tennessen, J. A., Kenny, E. E., Gravel, S., Mcgee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., and Boerwinkle, E. (2012), "Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes," *Science*, 64.

Tobacco, T., and Consortium, G. (2010), "Genome-wide meta-analyses identify multiple loci associated with smoking behavior," *Nature Genet.*, 42(5), 441–447.

Tzeng, J.-Y., and Zhang, D. (2007), "Haplotype-based association analysis via variance-components score test.," *Am. J. Hum. Genet.*, 81(5), 927–38.

Wakefield, J. (2009), "Bayes factors for genome-wide association studies: comparison with P-values)," *Genet. Epidemiol.*, 33, 79–86.

Waters, K. M., Stram, D. O., Hassanein, M. T., Le Marchand, L., Wilkens, L. R., Maskarinec, G., Monroe, K. R., Kolonel, L. N., Altshuler, D., Henderson, B. E., and Haiman, C. a. (2010), "Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups.," *PLoS Genet.*, 6(8).

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.," *Am. J. Hum. Genet.*, 89(1), 82–93.

Zhang, L., Yuan, F., Liu, P., Fei, L., Huang, Y., Xu, L., Hao, L., Qiu, X., Le, Y., Yang, X., Xu, W., Huang, X., Ye, M., Zhou, J., Lian, J., and Duan, S. (2013), "Association between PCSK9 and LDLR gene polymorphisms with coronary heart disease: case-control study and meta-analysis.," *Clin Biochem*, 46(9), 727–732.