A more powerful quasi-likelihood score test for detecting genetic association with multivariate phenotypes in related samples

Mingdong Liu

A thesis
submitted in partial fulfillment of
the requirements for the degree of

Master of Science

University of Washington

2015

Committee: Timothy Thornton, Chair

Kathleen Kerr

Alexander Reiner

Program Authorized to Offer Degree:

Public Health - Biostatistics

University of Washington

Abstract

A more powerful quasi-likelihood score test for detecting genetic association with multivariate phenotypes in related samples

Mingdong Liu

Chair of the Supervisory Committee

Dr. Timothy Thornton

Biostatistics

Pleiotropy is a commonly observed phenomenon in human genetics where a single gene influences multiple, and sometimes seemingly unrelated traits. Recently there has been significant interest in the identification of genetic variants that are associated with multiple phenotypes since identifying pleiotropic effects can lead to a better understanding of the underpinnings of complex traits. Genome-wide association studies often collect data on a variety of phenotypes, and a number of methods have been proposed for the joint analysis of multiple phenotypes in unrelated samples. Many genetic studies, however, include related individuals. In this thesis, we consider the problem of genetic association testing with multivariate phenotypes in samples with relatedness. We propose the multivariate phenotype quasi-likelihood (MPQ) score test for association mapping in related samples. Some of the features of the MPQ are: (1) it is applicable to completely general combinations of family and population-based samples, (2) it allows for the analysis of general quantitative traits and can accommodate both binary and continuous outcomes, (3) it can incorporate information on covariates in the analysis, and (4) it is computationally feasible for large-scale GWAS allowing for arbitrary relatedness among sample individuals. In simulation studies with unrelated and related samples, we demonstrate that the MPQ represents an overall, and in many cases, substantial, improvement, over existing multivariate methods, in terms of type-1 error rate and power, for a variety of causal models and multivariate trait correlation structures. Finally, we

apply the MPQ test to a GWAS of 3,548 Hispanic American postmenopausal women from the Women's Health Initiative SNP Health Association Resource to identify genetic variants associated with pleiotropic effects on serum C-reactive peptide (CRP) and white blood cell counts (WBC), two inflammation-related phenotypes. The MPQ test identifies previously reported variants for CRP and WBC as well as novel variants that are genome-wide significant.

Key words: genetic association testing, GWAS, multiple traits, quasi-likelihood score test, pleiotropy

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

Pleiotropy is a commonly observed phenomenon in human genetics where a single gene influences multiple, and sometimes seemingly unrelated traits. Recently there has been significant interest in the identification of genetic variants that are associated with multiple phenotypes since identifying pleiotropic effects can lead to a better understanding of the underpinnings of complex traits. Genome-wide association studies often collect data on a variety of phenotypes, and a number of methods have been proposed for the joint analysis of multiple phenotypes in unrelated samples. Many genetic studies, however, include related individuals. In this thesis, we consider the problem of genetic association testing with multivariate phenotypes in samples with relatedness. We propose the multivariate phenotype quasi-likelihood (MPQ) score test for association mapping in related samples. Some of the features of the MPQ are: (1) it is applicable to completely general combinations of family and population-based samples, (2) it allows for the analysis of general quantitative traits and can accommodate both binary and continuous outcomes, (3) it can incorporate information on covariates in the analysis, and (4) it is computationally feasible for large-scale GWAS allowing for arbitrary relatedness among sample individuals. In simulation studies with unrelated and related samples, we demonstrate that the MPQ represents an overall, and in many cases, substantial, improvement, over existing multivariate methods, in terms of type-1 error rate and power, for a variety of causal models and multivariate trait correlation structures. Finally, we apply the MPQ test to a GWAS of 3,548 Hispanic American postmenopausal women from the Women's Health Initiative SNP Health Association Resource to identify genetic variants associated with pleiotropic effects on serum C-reactive peptide (CRP) and white blood cell counts (WBC), two inflammation-related phenotypes. The MPQ test identifies previously reported variants for CRP and WBC as well as novel variants that are genome-wide significant.

Key words: genetic association testing, GWAS, multiple traits, quasi-likelihood score test, pleiotropy

Chapter 1

Introduction

The "common disease / common variant" hypothesis that common genetic variants in a population are largely responsible for common, complex diseases has been the basis for thousands of genome-wide association studies (GWAS) and candidate gene studies (Visscher, 2012). With the reduction of genotyping cost and the emergence of high-throughput technology, GWAS with millions of single nucleotide polymorphisms (SNPs) has become a widely used approach for detecting genetic variants that are associated with a variety traits and clinical outcomes (Jia, 2013). GWAS have successfully identified thousands of genetic variants that are associated with a variety of complex diseases and quantitative traits. Most genome-wide association studies perform single-variant association tests, where each variant in a genome-screen is tested individually for association with a single phenotype (See review: Sijde MR, 2014). An interesting observation from the large number of GWAS of complex traits that have been conducted is that genetic variants are often associated with multiple traits, where some of these traits are seemingly unrelated (Solovieff, 2013). This phenomenon can be caused in several ways. First, a gene can be pleiotropic (Stearns, 2010), that is, a single gene can directly control several phenotypes. For example, a mutation of one human tumor suppressor gene, p53, has been found to increase incidences of several types of tumors (Szymańska , 2003). Given that the human genome is complex, and there are millions of traits controlled by approximately 20,000 genes in the human genome, it is inevitable that a single gene can influence multiple traits. Pleiotropy is quite common in the human genome, and it has been estimated that approximately 16.9% of human genes are pleiotropic (Sivakumaran, 2011). Second, multiple phenotypes can be controlled by a single gene through regulatory networks. This phenomenon is believed to be more common than purely pleiotropic genes. In type 1 diabetes, for example, characteristically high blood glucose can cause both blurred vision and kidney disease (Weil, 1968). Third, multiple phenotypes may be pathophysiological counterparts,

such as serum uric acid level and gout, or strongly correlated, such as serum calcium and phosphorus, or even in a causal pathway, such as LDL cholesterol level and myocardial infarction (Pikula, 2015).

Most statistical methods for the analysis of multivariate phenotypes can be broadly classified into two groups: regression-based methods and data-reduction methods. Many regression approaches that are used for the analysis of multivariate phenotypes can be viewed as extensions of univariate regression methods to the multi-dimensional phenotype setting (Yang, 2012), including generalized estimating equations (GEE) (Murabito, 2007), generalized linear mixed model (GLMM), and frailty model. Data reduction methods reduce the dimensionality of multiple correlated phenotypes by deriving a new phenotype that is a function of the phenotypes of interest, and this new phenotype is used for inference on pleiotropic effects. Different dimensionality reduction criteria have been proposed including maximizing the variance explained for multiple phenotypes using principle component analysis (PCA) (Karasik, 2012), maximizing the total heritability via principal component of heritability (PCH) (Ott, 1999), or maximizing the quantitative trait locus heritability (PCQH), which is the variance explained by the new phenotype that is attributed to the genetic locus being tested for pleiotropy (Klei, 2008).

Many genetic studies of complex traits include related individuals. It is well known that failure to appropriately account for correlated genotypes and phenotypes among relatives in an association analysis can lead to both an increase in type-I error rate and a loss in power. GEE has previously been proposed for genetic association testing with multiple phenotypes in known pedigrees, where families are the clusters in the analysis. The FBAT-GEE method has been proposed for association testing that is robust to population stratification with multiple phenotypes in family-based designs (Christoph, 2003). FBAT-GEE is a multivariate score test based on GEE and is as an extension of the univariate transmission disequilibrium test (TDT) (Spielman, 1993). An advantage of FBAT-GEE is that the method does not make distributional assumptions on the phenotypes. FBAT-GEE, however, is restrictive as it requires genotype data on relatives, which may not be available for some individuals, and the method is not able to incorporate data on unrelated individuals, which can results in a substantial loss in power. Zeng

(Zeng , 2014) recently proposed the quasi-likelihood score (QLS) test for association testing with multiple phenotypes in general samples with relatedness. The QLS is a retrospective approach that models genotype as the outcome and the multivariate phenotypes as predictors. Kinship coefficients are used in the variance calculation of the score statistics for the QLS to account for the correlated genotypes among relatives.

In this thesis, we consider the problem of genetic association testing with multiple phenotypes in general samples with related individual. We develop the multivariate phenotypes quasi-likelihood score (MPQ) test. The MPQ test makes full use of the relationship information by explicitly modeling the enrichment effect of complex traits, which can be summarized as phenotypic values of an individual's relatives provides additional information, beyond that provided by the individual's own phenotypic value, about the probability that the individual carries an allele affecting the trait. Unlike the QLS, the MPQ can also incorporate information on covariates in the analysis. The MPQ method is computationally feasible for large-scale GWAS allowing for arbitrary relatedness among sample individuals. In simulation studies with related samples and a variety of causal models and multivariate trait correlation structures, we demonstrate that the MPQ represents an overall, and in many cases, substantial, improvement over existing multivariate methods for related samples, in terms of type 1 error and power, including GEE and the QLS method. Finally, we apply the MPQ test to a GWAS of 3,548 Hispanic American postmenopausal women from the Women's Health Initiative SNP Health Association Resource (WHI-SHARe) to identify genetic variants associated with pleiotropic effects on serum C-reactive peptide (CRP) and white blood cell counts (WBC), two inflammation-related phenotypes. The MPQ test identifies previously reported variants for CRP and WBC as well as novel variants that are genome-wide significant.

Chapter 2

Quasi-likelihood score test and MPQ test

2.1. Quasi-likelihood function

Let z_i $(i = 1, \dots n)$ be a random variable with expectation μ_i and variance V(μ_i), where V(μ_i) is some known function of μ_i. Suppose that μ_i is some function of $\boldsymbol{k}$ predictors $\boldsymbol{y_i} = (\boldsymbol{y_{i1}}, \boldsymbol{y_{i2}}, \dots, \boldsymbol{y_{ik}})$ and $\boldsymbol{k}$ +1 parameters $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ and

$$\boldsymbol{\mu_i = E(z_i) = f(\beta, y_i)} \qquad (1)$$

The quasi-likelihood function (Wedderburn, 1974), $U(z_i, \mu_i)$, is defined as

$$\frac{\boldsymbol{\partial U(z_i,\mu_i)}}{\boldsymbol{\partial \mu_i}} = \frac{\boldsymbol{z_i - \mu_i}}{\boldsymbol{V(\mu_i)}} \qquad (2)$$

Equivalently,

$$\boldsymbol{U(z_i, \mu_i) = \int^{\mu_i} \frac{z_i - \mu_i'}{V(\mu_i')} d\mu_i' + function\ of\ z_i} \qquad (3)$$

In practice, it is usually unnecessary to resolve quasi-likelihood function but the following quasi-likelihood properties are important:

$$\mathbf{E}\left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\mu_i}}\right) = \mathbf{0} \qquad (4)$$

$$\mathbf{E}\left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta_i}}\right) = \mathbf{0} \qquad (5)$$

$$\mathbf{E}\left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\mu}}\right)^2 = -\mathbf{E}\left(\frac{\partial^2 \mathbf{U}}{\partial \boldsymbol{\mu}^2}\right) = \frac{\mathbf{1}}{\mathbf{V(\mu)}} \qquad (6)$$

$$\mathbf{E}\left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta_i}} \frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta_j}}\right) = -\mathbf{E}\left(\frac{\partial^2 \mathbf{U}}{\partial \boldsymbol{\beta_i} \partial \boldsymbol{\beta_j}}\right) = \frac{\mathbf{1}}{\mathbf{V(\mu)}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta_i}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta_j}} \qquad (7)$$

Let

$$\mathbf{D} = \left[\frac{\boldsymbol{\partial U}}{\boldsymbol{\partial \beta_0}}, \frac{\boldsymbol{\partial U}}{\boldsymbol{\partial \beta_1}}, \dots, \frac{\boldsymbol{\partial U}}{\boldsymbol{\partial \beta_k}}\right] \qquad (8)$$

be a vector of derivatives of quasi-likelihood function with respect to the k parameters. The quasi-likelihood score function is defined as

$$\boldsymbol{U_{\beta_i}} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta_i}} \boldsymbol{V}(\boldsymbol{\mu})^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) \quad (9)$$

$$\mathbf{U} = \left[\boldsymbol{U_{\beta_0}},\ \boldsymbol{U_{\beta_1}}, \ldots \boldsymbol{U_{\beta_k}}\right] = \boldsymbol{D}'\boldsymbol{V}(\boldsymbol{\mu})^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \quad (10)$$

Since

$$\boldsymbol{Var}(\boldsymbol{U_{\beta_i}}) = \boldsymbol{Var}\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta_i}} \boldsymbol{V}(\boldsymbol{\mu})^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i})\right) = (\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta_i}})' \boldsymbol{V}(\boldsymbol{\mu})^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta_i}} \quad (11)$$

Similarly,

$$\boldsymbol{Cov}\left(\boldsymbol{U_{\beta_i}}, \boldsymbol{U_{\beta_j}}\right) = (\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta_i}})' \boldsymbol{V}(\boldsymbol{\mu})^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta_j}} \quad (12)$$

Put together, we have

$$\boldsymbol{Cov}(\boldsymbol{U}) = \boldsymbol{cov}\left[\boldsymbol{U_{\beta_0}},\ \boldsymbol{U_{\beta_1}}, \ldots, \boldsymbol{U_{\beta_k}}\right]$$

$$= \begin{bmatrix} \boldsymbol{Var}(\boldsymbol{U_{\beta_o}}) & \ldots & \boldsymbol{Cov}(\boldsymbol{U_{\beta_o}}, \boldsymbol{U_{\beta_k}}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{Cov}(\boldsymbol{U_{\beta_k}}, \boldsymbol{U_{\beta_0}}) & \ldots & \boldsymbol{Var}(\boldsymbol{U_{\beta_k}}) \end{bmatrix}$$

$$= \boldsymbol{D}'\boldsymbol{V}(\boldsymbol{\mu})^{-1}\boldsymbol{D} \quad (13)$$

2.2. General QLS test in genetic association study

Suppose we observe a bi-allelic marker and k traits in a sample with n individuals, where sample individuals can be arbitrarily related. Let

$$\boldsymbol{Y} = \begin{bmatrix} \mathbf{1} & \boldsymbol{y_{11}} & \cdots & \boldsymbol{y_{1k}} \\ \mathbf{1} & \boldsymbol{y_{21}} & \cdots & \boldsymbol{y_{2k}} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{1} & \boldsymbol{y_{n1}} & \cdots & \boldsymbol{y_{nk}} \end{bmatrix}$$

where y_{ij} is the trait j from individual i, and $y_i = (1, y_{i1}, y_{i2} \ldots y_{ik})$ is the ith row of Y for individual i. Assume the pedigree structure is available so that the inbreeding coefficients and kinship coefficients between individuals are also known. Suppose that for each observation, $\mu = (\mu_1,\ \mu_2, \ldots \mu_n)'$ is some functions of k parameters $\beta = (\beta_1,\ \beta_2, \ldots \beta_k)'$, and $G = (g_1, g_2, \ldots g_n)'$ represents the vector of genotypes at the marker for the sample individuals, with $g_i = 0, 1, 2$, representing wild type homozygotes, heterozygotes and variant homozygotes,

respectively. A retrospective regression method is used, that is, the genotype of the marker is treated as the outcome and the traits are treated as the predictors. Let $\mu = E(G|Y) = (\mu_1, \mu_2, \dots \mu_n)'$, with a logistic link, that is,

$$\boldsymbol{\mu_i = E(g_i|y_i) = \frac{\exp(y_i'\beta)}{1+\exp(y_i'\beta)}} \qquad (14)$$

If the marker is not associated with any of the traits, all coefficients in $(\beta_1, \beta_2, \dots \beta_k)'$ should be 0, and the appropriate hypotheses for testing that at least one of the traits is associated with the genetic marker is

$$\boldsymbol{H_o: \beta_1 = \beta_2 = \cdots = \beta_k = 0} \text{ versus } \boldsymbol{H_a: at\ least\ one\ \beta \neq 0}$$

Under an assumption that the pedigree founders in a sample are drawn from the population in Hardy-Weinberg equilibrium (HWE) for the given marker, the variance of the genotype can be defined through the kinship and the frequency of the genotype, as described by Bourgain (Bourgain, 2003). A general multivariate quasi-likelihood score function (QLS) proposed by Feng (Feng, 2014) has the form

$$\boldsymbol{U = \left[U_{\beta_0}, U_{\beta_1} \dots U_{\beta_k}\right]' = D'\Sigma^{-1}(G - \mu)} \qquad (15)$$

Where

$$\boldsymbol{D = \left[\frac{\partial \mu}{\partial \beta_o}, \frac{\partial \mu}{\partial \beta_1} \dots \frac{\partial \mu}{\partial \beta_k}\right]'} \qquad (16)$$

Σ is the covariance matrix of Y. Under the null hypothesis, all parameters are 0 except for $\boldsymbol{\beta_0}$, which is a nuisance parameter. We obtain the expectation of $\boldsymbol{\mu}$ under the null hypothesis by setting U_{β_0} equal to 0 and solving for $\boldsymbol{\beta_0}$. The estimated allelic frequency of μ under the null, $\hat{\mu}$, can be obtained from the estimated $\boldsymbol{\beta_0}$.

Under null hypothesis that traits are not associated with the genetic marker and the covariance of the marker reduces to Σ_o, which has the following form

$$\boldsymbol{\Sigma = \Sigma_o = Cov(G) = \frac{1}{2}\hat{\mu}(1 - \hat{\mu})\rho} \qquad (18)$$

$$\boldsymbol{\rho} = \begin{bmatrix} \mathbf{1}+\boldsymbol{\phi_1} & \mathbf{2}\boldsymbol{\phi_{12}} & \cdots & \mathbf{2}\boldsymbol{\phi_{1n}} \\ \mathbf{2}\boldsymbol{\phi_{21}} & \mathbf{1}+\boldsymbol{\phi_2} & \cdots & \mathbf{2}\boldsymbol{\phi_{2n}} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{2}\boldsymbol{\phi_{n1}} & \mathbf{2}\boldsymbol{\phi_{n2}} & \cdots & \mathbf{1}+\boldsymbol{\phi_n} \end{bmatrix} \quad (19)$$

where ρ is the correlation matrix, ϕ_i is the inbreeding coefficient of individual i, and ϕ_{ij} is the kinship coefficient between individual i and j.

With $\hat{\mu}$ and Σ_o, the regular quasi-likelihood score statistic for $(\beta_1, \beta_2 \dots \beta_k)$ proposed by Feng (2014) is given by

$$\boldsymbol{W = U^T_{-\beta_o} \Sigma^{-1}_{-\beta_o} U_{-\beta_o}} \quad (20)$$

where $U_{-\beta_o} = \left[U_{\beta_1} \dots U_{\beta_k} \right]'$ is a vector of score function similar to (15) except that U_{β_0} is omitted, and $\Sigma^{-1}_{-\beta_o}$ is a $k \times k$ matrix from the inverse matrix of Σ_o after the first column and the first row are removed. Under null hypothesis, W asymptotically follows a χ^2 distribution with $k-1$ degree of freedom.

2.3. MPQ test

In the general multivariate QLS test, correlated genotypes among sample individuals are accounted for, thus allowing for valid genetic association testing. However, dependences among trait values for sample individuals are ignored, which is a reasonable model if sample individuals are unrelated with unique environmental factors acting on the traits. However, when samples are drawn from families or there is a population with substructure, an assumption that traits are uncorrelated may not be appropriate. Traits of individuals from the same family are often correlated due to relatives having similar genetic background and sharing alleles identical by decent, and this correlation can provide additional information about the relationship between genetic marker and trait values. Jakobsdottir and McPeek (2013) proposed using kinship coefficients to model the enrichment effect of complex traits in related samples, which can be summarized as phenotypic values of an individual's relatives provides additional information, beyond that provided by the individual's own phenotypic value, about the probability that the individual carries an allele affecting the trait.

Let y_{ij} be the previously defined trait j from individual i. The enrichment phenotype, a_{ij}, can be defined as

$$\boldsymbol{a_{ij}} = \sum_{\boldsymbol{m=1}}^{\boldsymbol{n}} \emptyset_{\boldsymbol{mi}} \left(\boldsymbol{y_{mj}} - \widehat{\boldsymbol{y}}_{\boldsymbol{ij}}\right) \qquad (21)$$

where $\hat{y}_{ij}$ is an estimated mean of trait j for individual i, which can be a sample mean of the trait or a predicted value from a linear regression model that includes relevant covariates for the trait (but does not include the genetic marker being tested for association). The transformed phenotype for an individual is a weighted sum of mean adjusted trait values for individuals in the sample with weights proportional to the kinship coefficient among individuals.

Let

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{1} & \boldsymbol{a_{11}} & \cdots & \boldsymbol{a_{1k}} \\ \mathbf{1} & \boldsymbol{a_{21}} & \cdots & \boldsymbol{a_{2k}} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{1} & \boldsymbol{a_{n1}} & \cdots & \boldsymbol{a_{nk}} \end{bmatrix} \qquad (22)$$

be the matrix of enrichment phenotypes where $a_i = (1, a_{i1}, a_{i2} \ldots a_{ik})$ is the ith row of A and corresponds to the transformed phenotypes for individual i. Here we also use retrospective regression method as in general QLS test. The mean model that we are proposing takes into account the enrichment effect is

$$\boldsymbol{\mu_i} = \boldsymbol{E}(\boldsymbol{g_i}|\boldsymbol{a_i}) = \frac{\exp(\boldsymbol{a_i}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{a_i}'\boldsymbol{\beta})} \qquad (23)$$

Under null hypothesis, the quasi-score function for the mean model above is

$$\boldsymbol{U'} = \left[\boldsymbol{U'_{\beta_0}}, \boldsymbol{U'_{\beta_1}}, \ldots \boldsymbol{U'_{\beta_k}}\right]' = \boldsymbol{A}\boldsymbol{\Sigma}^{-1}(\boldsymbol{G} - \widehat{\boldsymbol{\mu}}) \qquad (24)$$

Comparing both statistic in general quasi-likelihood score test and MPQ, the difference between U' and U is the covariance term is removed in U' except for the centered term in U'. The statistic for MPQ, W_M becomes

$$\boldsymbol{W_M} = \boldsymbol{U'_{-\beta_o}}'\boldsymbol{\Sigma}^{-1}_{-\boldsymbol{\beta_o}}\boldsymbol{U'_{-\beta_o}} \qquad (25)$$

where

$$\boldsymbol{U'_{-\beta_o}} = \left[\boldsymbol{U'_{\beta_1}}, \boldsymbol{U'_{\beta_2}}, \ldots \boldsymbol{U'_{\beta_k}}\right]' \qquad (26)$$

Under null hypothesis, W_M asymptotically follows a χ^2 distribution with $k-1$ degree of freedom.

To improve power, we recommend using the predicted values of y_{ij} from a mixed linear model (MLM) when relevant covariates on sample individuals are available for the calculation of the enrichment phenotype given in equation 21. Thus, the term, $\emptyset_{mi}(y_{mj}-\hat{y}_{ij})$ in the calculation can be regarded as the residual of the trait j from individual i that is extracted from individual m and weighted on their kinship. $\hat{Y}_i$ is set to be a vector of estimation of traits of individual i from a mixed linear model. In MLM, the trait is treated as outcome, and genetic and environmental backgrounds are treated as random effects. Let X_i be a vector of covariates from individual i, such as age, race, etc., $g_i=(g_{i1},g_{i2}\dots g_{ik})$ is the vector representing the polygenic effects of individual i, $\varepsilon_i=(\varepsilon_{i1},\varepsilon_{i2}\dots\varepsilon_{ik})$ is the vector representing the environmental effects of individual i, where both g_{ij} and ε_{ij} are treated as random effects. As shown by Jakobsdottir and McPeek (Jakobsdottir, 2013), the predicted values of the traits can be obtained using the following model:

$$\boldsymbol{y_{ij}=X_i\beta_i+g_{ij}+\varepsilon_{ij}}$$

$$\boldsymbol{g_i}\sim\boldsymbol{MVN}[\mathbf{0},\boldsymbol{\sigma_a^2}\boldsymbol{\rho}]$$

$$\boldsymbol{\varepsilon_i}\sim\boldsymbol{MVN}[\mathbf{0},\boldsymbol{\sigma_e^2}\mathbf{I}]$$

Chapter 3

Data simulation and real data

3.1. General simulation setting

We consider the setting where there are three traits of interest, (y_1, y_2, y_3), that may be correlated or uncorrelated, with a di-alleilic casual gene (D/d) that is linked with a genetic marker (A/a) to be tested for association. y_1 is defined as a trait that can be on causal pathway from the genotype. y_2 is defined as a trait that can be on the causal pathway from the genotype and/or y_1. Similarly, y_3 is defined as the trait that can be on the causal pathway from any or all of genotype, y_1, and y_2. Traits of y_1, y_2, and y_3 can also be completely independent of genotype, which correspond to null models of association. That trait models considered can be written as

$$\boldsymbol{y_1 = \beta_1 G + \varepsilon_1}$$

$$\boldsymbol{y_2 = \beta_2 G + \gamma_{12} y_1 + \varepsilon_2}$$

$$\boldsymbol{y_3 = \beta_3 G + \gamma_{13} y_1 + \gamma_{23} y_2 + \varepsilon_3}$$

For simplicity, we assume $\varepsilon_j \sim N(0,1)$, for $j = 1, 2, 3$. To allow for other extraneous variables that may cause correlation of 3 traits, we also include correlated random effects for three traits where $(\varepsilon_1, \varepsilon_2, \varepsilon_3)' \sim MVN(0, \Sigma)$. In our simulation studies, we set $\Sigma_{ij} = 0.2$, $for\ i \neq j$.

Heritability and interability are used to generate the regression coefficients. The heritability, h^2, is defined as the proportion of phenotypic variation that can be explained by the variation of genetic effects. Let p be the frequency of reference allele at the genetic marker that is to be tested for association with the traits. As shown by Zhu et al. (Zhu, 2009), we define

$$\boldsymbol{h_1^2 = \frac{var(\beta_1 G)}{var(y_1)} = \frac{2\beta_1^2 p(1-p)}{1 + 2\beta_1^2 p(1-p)}}$$

Thus,

$$\boldsymbol{\beta_1 = \frac{1}{\sqrt{2p(1-p)}} \sqrt{\frac{h_1^2}{1 - h_1^2}}}$$

Similarly,

$$\boldsymbol{\beta_2} = \frac{\mathbf{1}}{\sqrt{\mathbf{2p(1-p)}}}\sqrt{\frac{\boldsymbol{h_2^2}}{\mathbf{1}-\boldsymbol{h_2^2}-\boldsymbol{\gamma_{12}^2}}}$$

$$\boldsymbol{\beta_3} = \frac{\mathbf{1}}{\sqrt{\mathbf{2p(1-p)}}}\sqrt{\frac{\boldsymbol{h_3^2}}{\mathbf{1}-\boldsymbol{h_3^2}-\boldsymbol{\gamma_{13}^2}-\boldsymbol{\gamma_{23}^2}}}$$

Interability, k_{ij}^2, is defined as the contribution of variation of trait i to the variation of trait j, for example,

$$\boldsymbol{k_{12}^2} = \frac{\boldsymbol{var(\gamma_{12}y_1)}}{\boldsymbol{var(y_2)}} = \frac{\boldsymbol{\gamma_{12}^2(1-h_2^2)}}{\mathbf{1}+\boldsymbol{\gamma_{12}^2}}$$

Solve for $\boldsymbol{\gamma_{12}}$, we get

$$\boldsymbol{\gamma_{12}} = \sqrt{\frac{\boldsymbol{k_{12}^2}}{\mathbf{1}-\boldsymbol{h_2^2}-\boldsymbol{k_{12}^2}}}$$

Similarly,

$$\boldsymbol{\gamma_{13}} = \sqrt{\frac{\boldsymbol{k_{13}^2}}{\mathbf{1}-\boldsymbol{h_3^2}-\boldsymbol{k_{13}^2}-\boldsymbol{k_{23}^2}}}$$

$$\boldsymbol{\gamma_{23}} = \sqrt{\frac{\boldsymbol{k_{23}^2}}{\mathbf{1}-\boldsymbol{h_3^2}-\boldsymbol{k_{13}^2}-\boldsymbol{k_{23}^2}}}$$

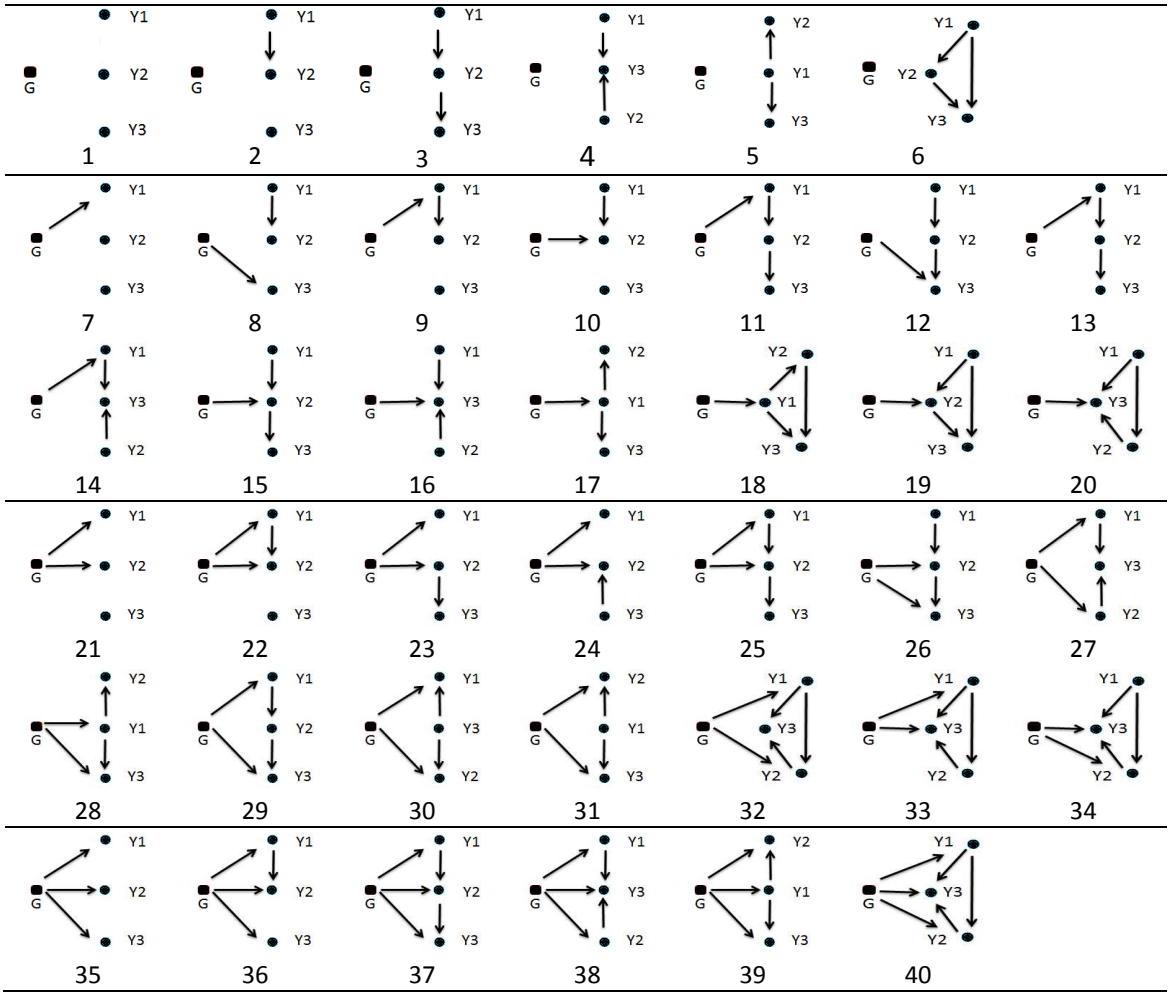
Therefore, we can derive all βs and γs once heritability and interability are specified. Heritability is set to the same value in our simulation studies, and we consider traits where $h^2 = 0.05$. This is weak heritability as compared to some highly hereditary traits, such as blood pressure with $h^2 > 0.3$ (Miall, 1963). We consider three interability levels: 0.05, 0.15 and 0.35.

In order to examine the impact of phenotypic interaction on the type I error rates and power, we construct 39 causal models. These models are illustrated with directed acyclic graphs (DAGs) (**Figure 3.1**), as described by Zhu (Zhu, 2009). In Models 2-6, the genetic marker is not associated with either trait, thus they can be used to evaluate type I error. In models 7-20, the genetic marker is associated with only one trait. In models 21-34, the genetic marker is

associated with 2 traits. In models 35-40, the genetic marker is associated with all 3 traits. All of the models that have at least on trait that is associated with the genetic marker are used to estimate the power.

For each model, 5000 simulated replicates are obtained to calculate empirical type I error and power.

Figure 3.1. DAGs of simulated models



3.2. Unrelated Samples

In the simulation studies with unrelated individuals, sample sizes of 300, 450 and 600 are considered. All traits are continuous variables.

3.3. Pedigree data

The simulated pedigree data contains 100 unrelated individuals and 10 families. Each family contains 16 individuals. The genogram is shown in **Figure 3.2**. Assignment of genotype in family members (6-8 and 9-16) complies with HWE.

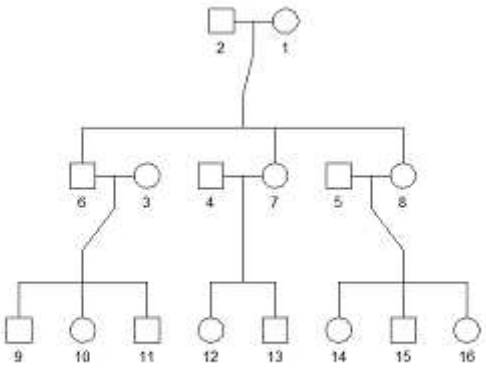


Figure 3.2. The genogram for simulation of pedigree data.

The simulation of unrelated individuals is similar to what is described in section 3.2. Two sets of trait variables are considered. In the first setting, all 3 traits are continuous variables, as described in section 3.1. In the second setting, we include a dichotomous trait. We simulate 3 continuous traits, and we apply a threshold to the first trait to generate a binary variable. We choose mean + 1.1*standard error so that the proportion of the case is around 12% for the binary trait. The other two traits in setting 2 are continuous variables and are not transformed.

3.4. Statistical tests evaluated in the simulation studies

To compare the performance of MPQ test, several statistical tests are evaluated and compared using the simulated data. These competing tests are described in this section. For all tests, we set the type I error to be 0.01. In the following models, the 3 traits are referred as y_1, y_2, and y_3, and the genotype of the marker is referred as g, as in the previous chapter.

Retrospective GLM

The retrospective logistic GLM method is used with the genotype as the outcome. Let μ be the frequency of the reference allele at the genetic maker to be tested for association with the phenotypes, The GLM model used is

$$\boldsymbol{logit(\mu_i) = \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2} + \beta_3 y_{i3} + \varepsilon_i}$$

where μ_i is the expected marker genotype value of subject i, $y_{1i}, y_{2i}, and\ y_{3i}$ are the 3 traits of subject i. The null hypothesis is all βs are zero versus the alternate that at least one β is not zero.

Simple linear regression model

In this prospective simple linear regression model, each trait is individually regressed on the genotype.

$$\boldsymbol{y_{ij} = \beta_{0j} + \beta_{1j} g_i + \varepsilon_i}$$

y_{ij} is the trait j $(j = 1, 2, 3)$ of subject i $(i = 1, \dots n)$, and g_i is the marker genotype. The null hypothesis of each individual test is $\beta_j = 0$. To control overall type I error of multiple tests, Bonferroni correction is applied. The significance level α of each individual test is set to be 0.01/3.

Principle component analysis (PCA)

For the PCA test, the 3 phenotypes are first reduced to a single variable using eigenvalue decomposition (Pearson, 1901),

$$\boldsymbol{\hat{y}_i = a y_i} \quad i = 1, 2, \dots n$$

where y_i is the vector of 3 traits of individual i, and a is the eigenvector of the covariance of Y with maximal eigenvalue. $\hat{y}_i$ is the reduced, one dimensional new variable that is regressed on the genotype.

$$\boldsymbol{\hat{y}_i = \alpha + \beta g_i + \varepsilon_i}$$

g_i is the genotype for individual i and ε_i is the error of regression model. Only one parameter, β, is to estimated and tested. The null hypothesis is $\beta = 0$.

Generalized Estimating equation model (GEE)

Let $y_i = (y_{i1}, y_{i2}, y_{i3})'$ be 3 traits of individual i, and $G_i = (g_i,\ g_i, g_i)'$, where g_i is the genotypes of individual i, $g_i = 0, 0.5, 1$ for normal homozygote, heterozygote, and homozygote of investigated allele, respectively. For samples with unrelated individuals, traits are clustered by individuals, and a working correlation matrix between traits from same subject is specified. For samples with pedigrees, clustering should be done at the family level (Liang, 1986).

$$\boldsymbol{y_i = G_i'\beta + \varepsilon_i}$$

General QLS test

General QLS test uses the same model as GLM,

$$\boldsymbol{g_i = \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2} + \beta_3 y_{i3} + \varepsilon_i}$$

g_i is the genotype of individual i, y_{i1}, y_{i2}, and y_{i3} are 3 traits of individual i, and a mean model of genotype is used instead of logit link.

3.5. Description of real data from the Women's Health Initiative

We applied our MPQ test to a real data from the Women's Health Initiative SNP Health Association Resource (WHI-SHARe). The study was conducted in 40 clinical centers in 24 states and the District of Columbia, and total 161,808 postmenopausal women between 50 and 79 years old were recruited into the study. Clinical information was collected by self-report, and physical and laboratory examination. Here we are interested in identifying genetic loci that are associated with two inflammatory phenotypes, white blood cells count and serum C-reactive peptide level, in Hispanic women. There are 3587 WHI-SHARe Hispanic participants, CRP and WBC was successfully measured on 3548 and 3551 Hispanics participants, respectively, and a total 3512 participants have both CRP and WBC measurements available.

The WBC phenotype was log(10)-transformed (logWBC), and a natural log-transformed was used for CRP (CRP) for the association analysis. Histograms of CRP and logWBC in 3512 participants are shown in **Figure 3.3**. The mean and median of CRP is 1.128 mg/L and 1.151 mg/L, respectively, and the interquartile range was 0.451–1.183. For logWBC, the mean and median is is 1.764 and 1.758, respectively, and the interquartile range is 1.609-1.932. The

scatterplot of CRP vs logWBC is given in **Figure 3.4**. The correlation between CRP and logWBC is 0.3635. The heritability of CRP was estimated to be 0.287 with a 95% CI of (0.139, 0.435), and the heritability of logWBC was estimated to be 0.407 with a 95% CI of (0.248, 0.565).

DNA was extracted by the Specimen Processing Laboratory at the Fred Hutchinson Cancer Research Center from specimens that were collected at the time of enrollment. Genome-wide genotyping was performed at Affymetrix on the Affymetrix 6.0. SNPs on the Y chromosome and Affymetrix quality-control (QC) probes were excluded from analysis. After exclusions and quality control filtering of SNPs, there are 829,370 genotyped SNPs.

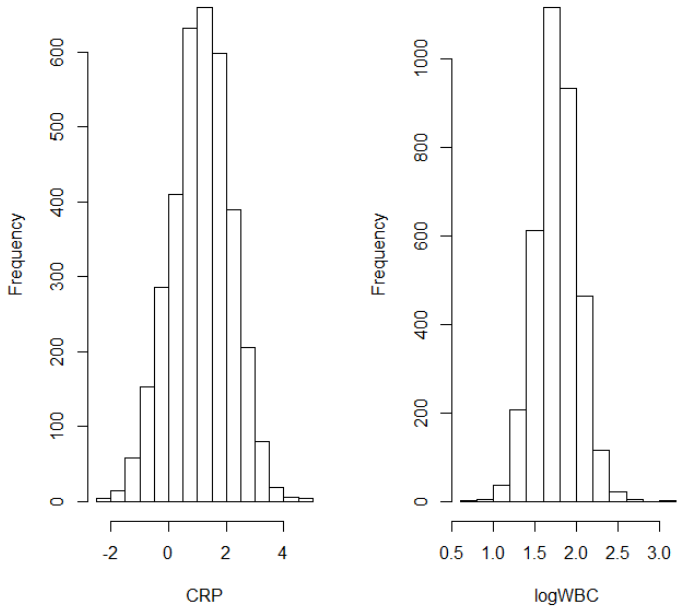


Figure 3.3. Histogram of CRP and logWBC of Hispanic participants.

3.6. Statistical test for real data

We first fit the data with a mixed linear model (MLM), age and race are treated as the fixed effects. Let $g_i = (g_{i1}, g_{i2} \dots g_{ik})$ be the vector representing the polygenic effects of individual i, $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2} \dots \varepsilon_{ik})$ is the vector representing the environmental effects of individual i, both g_{ij} and ε_{ij} are treated as random effects (Jakobsdottir, 2013),

$$\boldsymbol{y_{ij} = \beta_{0j} + \beta_{1j} * Age_i + \beta_{2j} * Race_i + g_{ij} + \varepsilon_{ij}}$$

$$\boldsymbol{g_i \sim MVN[0, \sigma_a^2 \rho]}$$

$$\boldsymbol{\varepsilon_i \sim MVN[0, \sigma_e^2 I]}$$

Based on this model, we can obtain the predicted values of the 3 traits, $(\hat{y}_{i1}, \hat{y}_{i2}, \hat{y}_{i3})$, for each individual i. The residuals, $y_i - \hat{y}_i$, were then used to obtain the enrichment phenotypes for the MPQ model. Since pedigree information is not available on the sample individuals, empirical kinship coefficients between all pairs of study participants were computed using the genome-screen data and a method-of-moments that accounts for ancestry admixture among sample individuals (Thornton, 2010).

We also considered the recently proposed Multivariate Outcome Score Test (MOST) by He et al. (He, 2013) to the WHI-SHARe Hispanic samples for joint association testing with CRP and WBC. MOST is based on GEE with clustering on the individual.

3.7. Software for simulation and statistical tests

We use R 3.1 for simulation and statistical testing, including the following R packages: MVTNORM, GEE, GWASTOOLS, QQMAN.

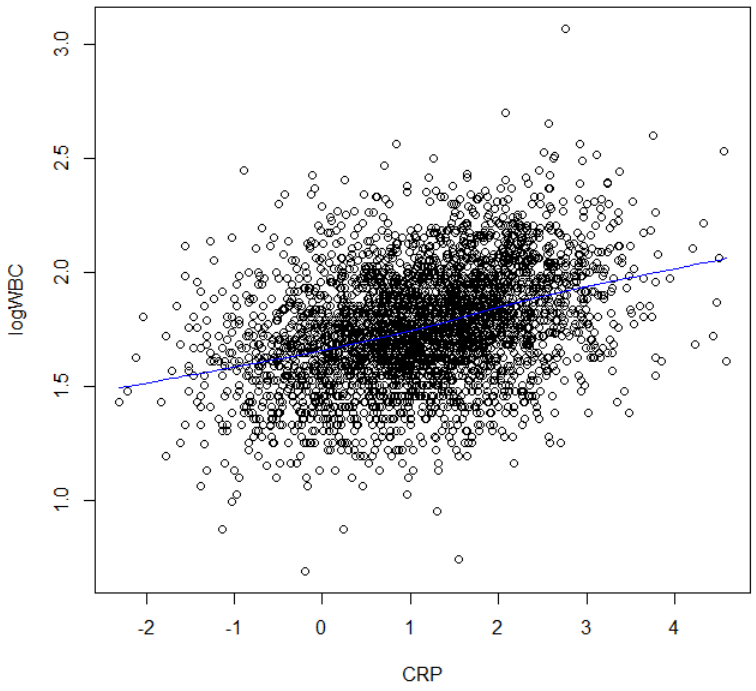


Figure 3.4. Scatterplot of CRP and logWBC. The fitted line is shown and the correlation coefficient between them is 0.3635.

Chatper 4

Results

4.1. Unrelated Samples

4.1.1. Type I error

In models 2-6, the 3 phenotypes are unrelated to the genotype, but these traits may either be correlated or uncorrelated. These 5 models are suitable for estimation of type I error, and the result are shown in **Table 4.1**.

Note that the MPQ test and general QLS test are equivalent tests for population-based samples with unrelated individuals when there are no covariates. The empirical type I errors from MPQ, LM, GLM and PCA models are not significantly different from the nominal level, and increasing interability level has no effect on type I error. In contrast, estimated type I error rates from GEE model are significantly higher than the nominal significance level in most cases (values that are significantly different from the nominal level are in bold).

Table 4.1. Estimated type I errors with population data (n=300, α=0.01).

Interability	Model No.	GLM	LM	MPQ	PCA	GEE
0.05	2	0.0092	0.0106	0.0106	0.0086	**0.0140**
	3	0.0108	0.0104	0.0120	0.0112	**0.0154**
	4	0.0100	0.0092	0.0102	0.0114	**0.0158**
	5	0.0052	0.0084	0.0066	0.0116	0.0104
	6	0.0076	0.0102	0.0076	0.0098	0.0124
0.15	2	0.0086	0.0086	0.0098	0.0104	**0.0142**
	3	0.0088	0.0114	0.0106	0.0082	**0.0160**
	4	0.0084	0.0078	0.0086	0.0094	0.0128
	5	0.0090	0.0076	0.0100	0.0070	**0.0156**
	6	0.0088	0.0098	0.0100	0.0108	**0.0146**
0.35	2	0.0068	0.0088	0.0074	0.0092	0.0118
	3	0.0076	0.0088	0.0088	0.0104	0.0130
	4	0.0078	0.0090	0.0086	0.0108	0.0120
	5	0.0108	0.0076	0.0122	0.0102	**0.0154**
	6	0.0108	0.0086	0.0130	0.0088	**0.0156**

4.1.2. Power

In models 7-40, there is at least one trait that is associated with the genetic marker, and these models were used to assess the power of the tests. For models 7-20, there is only one trait that is associated with the marker. Two traits are associated with the marker in models 21-34, and all 3 traits are associated with the marker in models 35-40.

Power results are given in **Figure 4.1** for samples with 300 unrelated individuals. Among the five statistical tests considered, the retrospective GLM and MPQ test have very similar power for all models under the three interability settings. This is not surprising because the two tests use the same model. When subjects are unrelated, the covariance matrix in formula (18) in section 2.2 reduces to that used in GLM model. In addition, the quasi-likelihood function is same as general likelihood function for the distribution of exponential family (Wedderburn, 1974). GLM and MPQ generally have the highest power, while simple linear regression with Bonferroni correction is always conservative except for the settings when the 3 traits are all associated with the genotype. The power of PCA test is comparable to GLM and MPQ only in the cases that all 3 traits are associated with the marker.

For all five statistical tests considered, there is a gain power when the genetic marker is associated with more than one trait. Power, however, is quite variable, even for models with same number of traits that are associated with the genetic marker. Consider models 7-20, for example, that have one trait directly associated with the genetic marker. We categorize the models based on empirical power levels into 3 groups. Group 1 has low power and consists of models 7, 9, 11, 13, 17 and 18. Group 2 has moderate power and includes models 8, 10, 14, 15, 16 and 19. Models 12 and 20 have high power and are the third group. The difference in power among the groups is caused by how the other two traits contribute to the association in the models for the groups, which can be explained through the DAGs. In group 1, the two traits are either independent of the associated trait, or lies in the causal pathway from the associated trait. They do not add any information to the association between the marker and the trait, however, the test power is impaired by increased degree of freedom. In group 2, the associated trait is dependent on only one of remaining, yet unassociated traits. Although such dependence does not add any information to the association between the marker and the trait, conditioning

on this unassociated trait can reduce the variance of the associated trait, therefore increase the test power. This unassociated trait is usually called precision variable. In group 3, the power is further improved because both traits play the role of precision variables.

The effect of interability change can also be explained by the model structures. Since the associated trait is independent of the other two traits in group 1, an increase in interability does not modify the association, and the power still remain unchanged with higher interability. As a contrast, when the interability increases, the effect of precision also becomes more evident, as observed by increased powers.

For each experimental setting, we also investigated how differences in sample size affects power. We considered sample sizes of 300, 450, and 600, and the power results for GLM and MPQ are given in **Figure 4.2**. As expected, power for all 4 tests increased as sample size increased. We also investigated how interability differences impacts power, and we estimated power of GLM and MPQ at interabilites values of 0.05, 0.15 and 0.35. As shown in Figure 4.2, the change of interability has little effect on the power of the tests for most models settings, except for the models where interability modifies the association, as discussed in the previous paragraph.

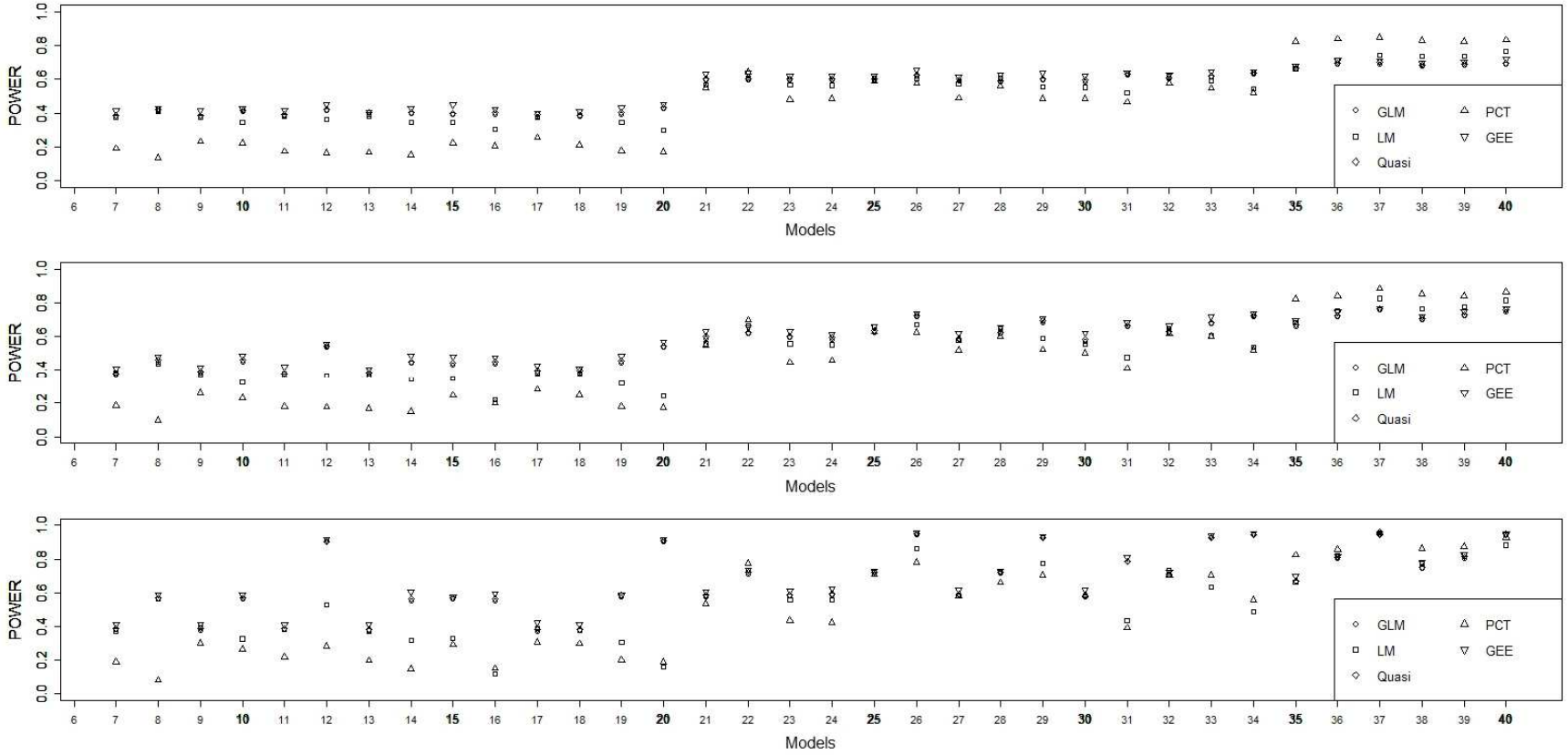


Figure 4.1. Statistical powers of GLM, MPQ, LM, PCA and GEE with unrelated data (n=300). Top: $k_{ij}^2 = 0.05$; Middle: $k_{ij}^2 = 0.15$; Bottom: $k_{ij}^2 = 0.35$.

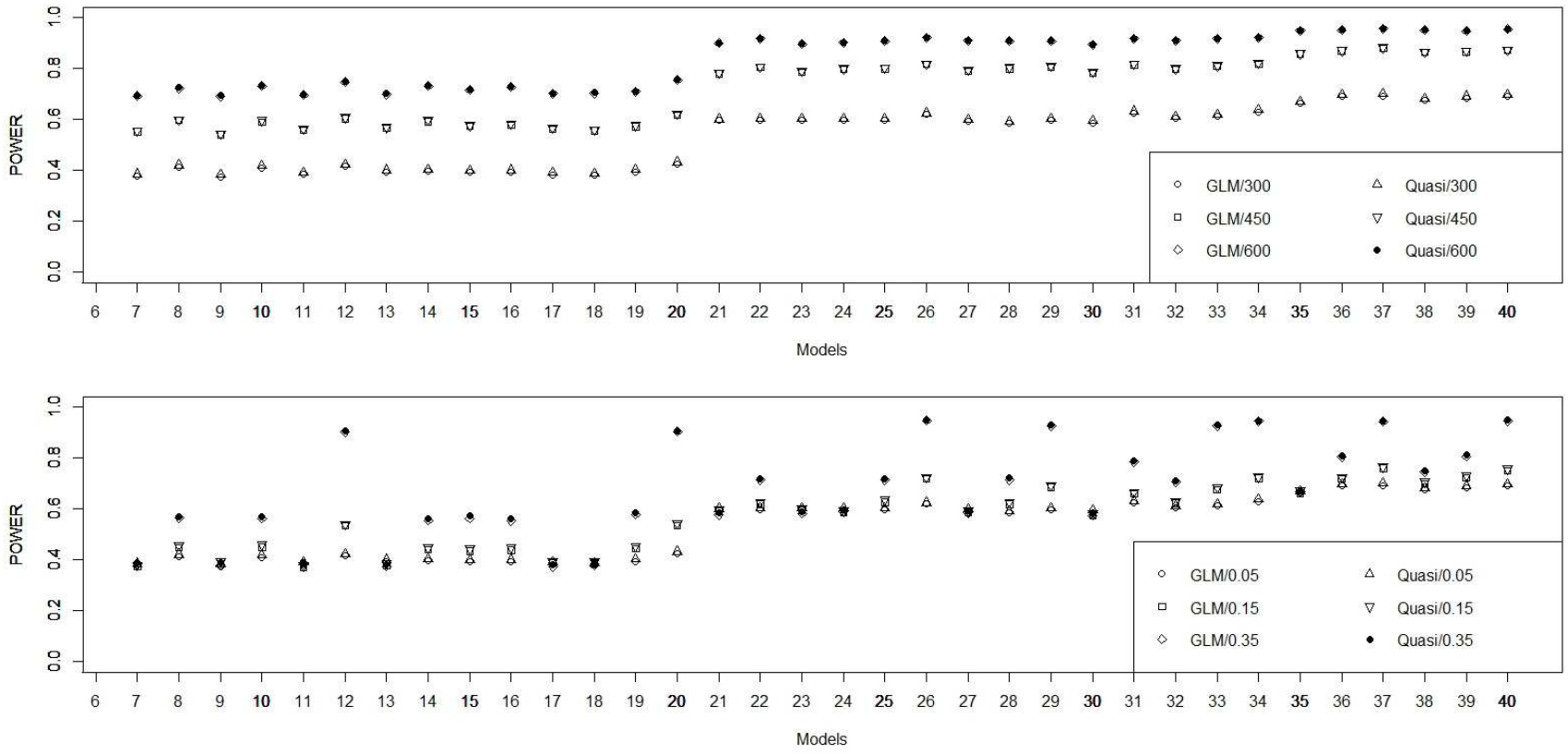


Figure 4.2. Assessing difference in statistical power of GLM and MPQ tests under different sample size and interabilities (k_{ij}^2) with unrelated samples. Upper: Power of GLM and MPQ tests for the dataset with 300, 450 and 600 individuals; Bottom: Power of GLM and MPQ tests for the dataset with interabilites of 0.05, 0.15 and 0.35.

4.2. Unrelateds and Pedigree data

4.2.1. Type 1 error

We conducted simulation study to assess type 1 error of the MPQ test in samples with 100 unrelated individuals and 10 pedigrees, as previously discussed. Note that GLM, PCA and LM approaches are not valid methods for association testing in related samples, so these approaches were not considered in this simulation study. We also compare empirical type 1 error of the MPQ test to the general multivariate QLS test of Feng (2014) and GEE with the family as the clustering unit. The empirical type I error rates of three tests are shown on **Table 4.2**. At the three interability levels considered, both MPQ test and general quasi-likelihood score test yield similar empirical type I errors that are not significantly different from the nominal level. In addition, these methods are properly calibrated when one of the traits is dichotomous and the other two traits are quantitative. GEE, however, has inflated type I error in most of the settings considered, and empirical type-1 error values with GEE that are significantly different from the nominal level are in bold font in the table.

4.2.2. Power

We did not include GEE in the evaluation of empirical power in the related samples since the method is not properly calibrated. Power results for the MPQ test and the general QLS are given in Figure 4.3. Similar to the results for the unrelated samples, as the number of traits that are associated with the genetic marker (**Figure 4.3** for continuous traits and **Figure 4.4** for combinative traits) increases, power for tests generally increases. For all settings considered, with either continuous traits or combinations of continuous and binary traits, the MPQ test generally has higher power than the regular QLS test. Interestingly, the empirical power of the MPQ test in the case when only one trait is associated with the marker is even higher than the empirical powers from regular QLS test in the case when two traits are associated with the marker. Surprisingly, interability has less effect on the power of MPQ than general QLS test. When the interability is small, general QLS test usually has lower powers, while the power of MPQ test is much more stable over interabilities. Compared to the general QLS test, the power

of the MPQ test is less affected by the type of trait model and type of model, and this is more evident in the simulation settings when all three traits are continuous (**Figure 4.5**).

Table 4.2. Empirical type I error rates under different interabilities (α=0.01)

k_{ij}^2	Model	Continuous traits only			Continuous and binary traits		
		Quasi	MPQ	GEE	Quasi	MPQ	GEE
0.05	2	0.009	0.0108	**0.0176**	0.0114	0.0112	**0.0156**
	3	0.0084	0.009	**0.0158**	0.0138	0.0122	**0.0144**
	4	0.0114	0.0106	**0.0154**	0.011	0.0094	**0.0148**
	5	0.01	0.0092	0.0126	0.0106	0.01	**0.016**
	6	0.0104	0.0088	**0.0132**	0.0092	0.0086	**0.0146**
0.15	2	0.0124	0.009	**0.0146**	0.0118	0.011	**0.0162**
	3	0.0102	0.0106	**0.0168**	0.0094	0.013	**0.0184**
	4	0.0112	0.0102	**0.0152**	0.0094	0.0082	0.0126
	5	0.012	0.009	**0.0134**	0.0106	0.01	**0.0164**
	6	0.0124	0.0122	**0.0164**	0.0112	0.0078	**0.0146**
0.35	2	0.0132	0.0084	**0.0166**	0.0092	0.0104	**0.0164**
	3	0.0126	0.0112	**0.0144**	0.0098	0.0082	**0.0142**
	4	0.0108	0.009	0.0128	0.011	0.0112	**0.0172**
	5	0.0106	0.0096	**0.0150**	0.0126	0.0114	**0.0176**
	6	0.0102	0.008	**0.0144**	0.0126	0.0112	**0.0166**

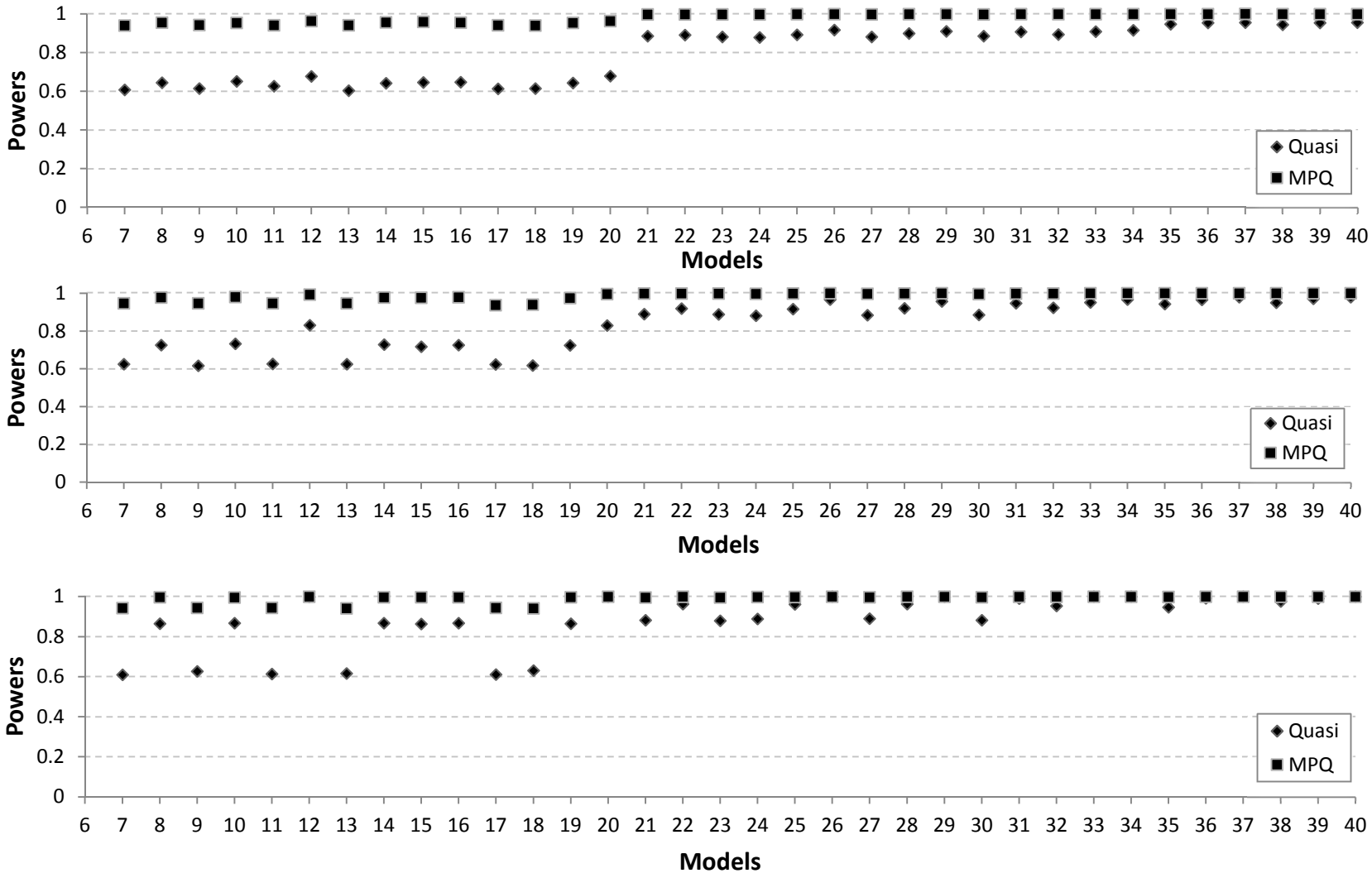


Figure 4.3. Powers of general QLS test and MPQ test with pedigree data containing only continuous traits . Upper: k_{ij}^2=0.05; Middle: k_{ij}^2=0.15; Bottom: k_{ij}^2=0.35.

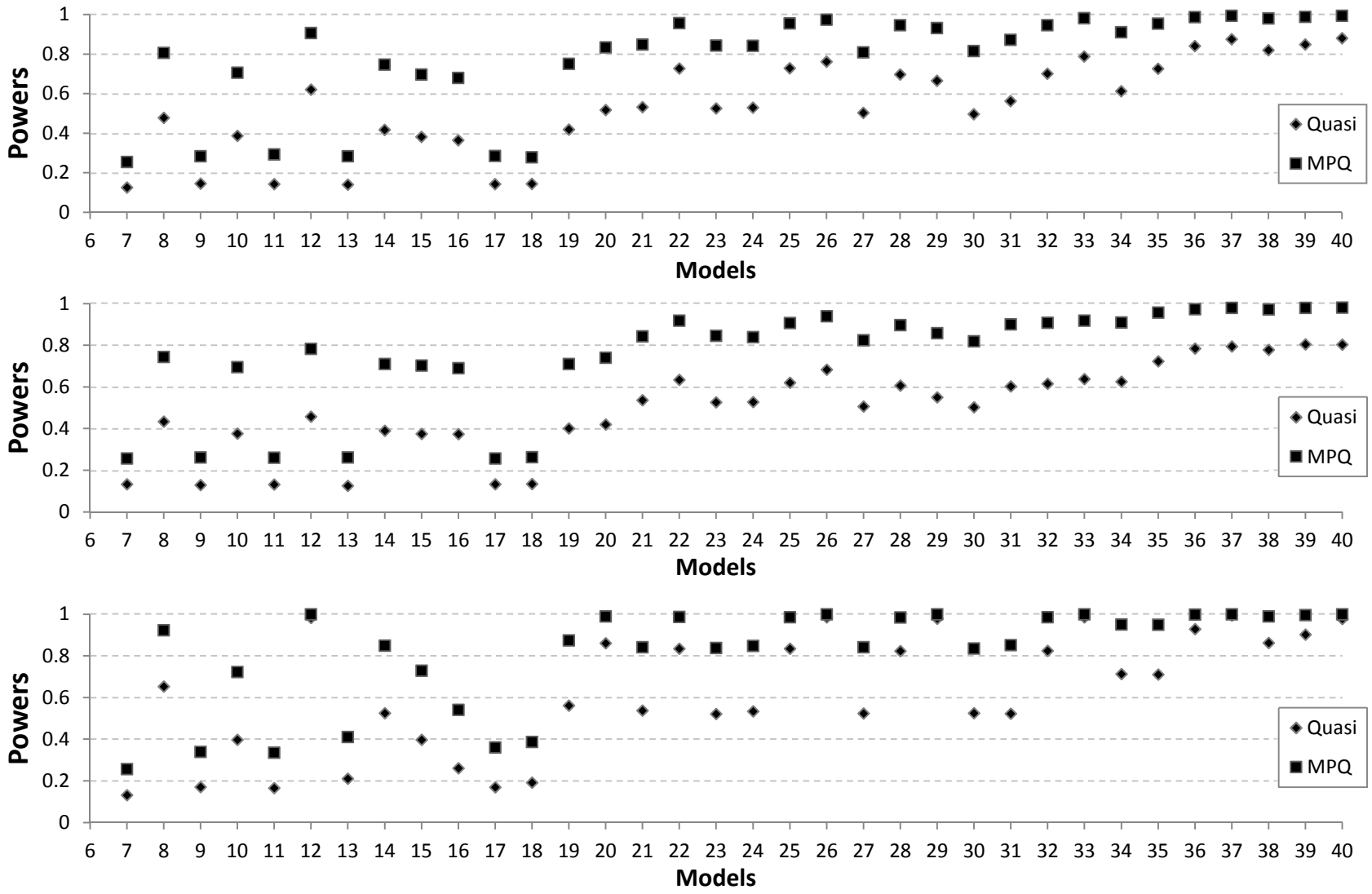


Figure 4.4. Powers of general QLS test and MPQ test with pedigree data containing continuous and binary traits. Upper: k_{ij}^2=0.05; Middle: k_{ij}^2=0.15; Bottom: k_{ij}^2=0.35.

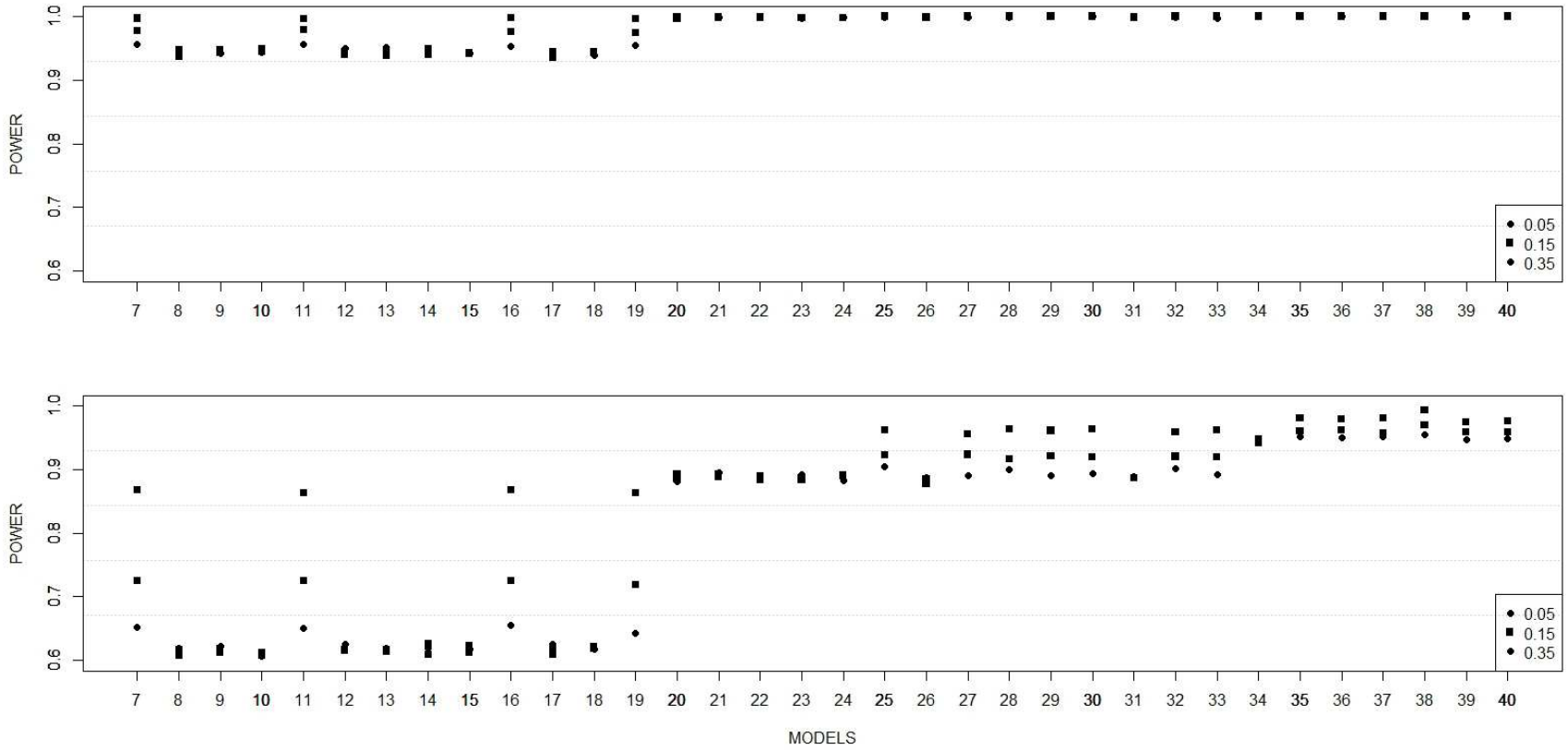


Figure 4.5. Empirical power of general QLS test and MPQ test over different interabilities on simulated samples with pedigrees and unrelated individuals. Upper: MPQ test; Bottom: regular quasi-likelihood score test.

4.3. Real data

We performed an association analysis with MPQ test to identify pleiotropic effects on CRP and logWBC in the WHI-SHARe Hisapnics. As a comparison, we also performed a univariate quasi-likelihood score test with CRP and with logWBC. A Manhattan plot of GWAS result is shown in **Figure 4.6**, and SNPs with significant p values and their genomic location are listed on **Table 4.3**. The genome-wide significant SNPs identified with the MPQ model are located in 3 genomic regions: 1p31 (chr1: 65874845-65929318), 1q23 (chr1: 156906658-157965173) and 12q24 (chr12: 119873345-119935808). The representative genes inside the loci are LEPR, CRP/DARC, and HNF1, respectively (**Figure 4.7**). Most SNPs identifed can also be detected in univariate models, however, SNPs identified by each univariate model are exclusive. For example, rs11265198, rs7534472, rs2808666, r856065, and rs857682 are only identified in the univariate model with logWBC but not with CRP. The genes within the 3 genomic regions with significant SNPs are plotted in **Figure 4.7**. Interestingly, two SNPs, rs12239267 and rs16827466, are detected only with the MPQ analysis but not with the univariate models.

Even though data are drawn from a general population that pedigree structure is not available, it is possible for us construct an empirical kinship from large amounts of SNPs sequenced, and this empirical kinship was used to perform the MPQ test.

We also applied the MOST method to the data and the same significant SNPs as MPQ were identified. MOST, however, was not properly calibrated, similar to what we observed in our simulations with GEE, with a genomic control inflation factor of 1.05. In contrast, MPQ is properly calibrated, genome wide, with a genomic control inflation factor of 0.98. MPQ test with empirical kinship is as powerful as GEE and has better control of type-1 error. The QQ plots for MOST and MPQ are shown in **Figure 4.8** where MOST has early deviation from the expected p-values under the null hypothesis compared to MPQ .

Table 4.3. SNPs detected with MPQ test.

SNP	chr	loc	MAF	Multivariate P value	LogWBC P value	CRP P value
rs2808629	1	157943420	0.347254005	7.32E-12	0.183948084	1.21E-09
rs2794520	1	157945440	0.345275478	8.57E-12	0.237364597	8.48E-10
rs7553007	1	157965173	0.344817768	2.05E-11	0.259224435	1.48E-09
rs11265198	1	157686931	0.077153451	.12E-11	5.29E-11	0.732029316
rs2259816	12	119919970	0.378667046	1.92E-10	0.686366145	1.88E-10
rs2393791	12	119908339	0.404443179	2.23E-10	0.729088403	2.42E-10
rs16827466	1	157916324	0.017084282	2.49E-10	0.000430702	6.74E-05
rs7534472	1	157737275	0.024094668	5.77E-10	3.18E-09	0.684629652
rs1169313	12	119927053	0.3809456	6.70E-10	0.692306494	5.97E-10
rs12239267	1	157937552	0.016900602	8.88E-10	0.000349526	0.000201283
rs7953249	12	119888107	0.418992027	3.71E-09	0.36930666	1.26E-09
rs2808666	1	157827940	0.028758542	5.37E-09	7.62E-10	0.007452042
rs2258043	12	119935808	0.467737913	9.01E-09	0.951191788	1.59E-08
rs1305096	1	65374345	0.450533379	1.13E-03	.0.206645699	2.44E-09
rs2258287	12	119938696	0.35759043	1.46E-08	0.439701154	6.16E-09
rs856055	1	157250067	0.022941009	1.94E-08	2.20E-08	0.910850266
rs1892534	1	65878532	0.454869021	2.24E-08	0.222053025	4.72E-09
rs857682	1	156906658	0.043304843	2.67E-08	3.76E-08	0.992606838
rs2650000	12	119873345	0.329441913	.66E-08	0.405387536	1.49E-08
rs4655772	1	65902986	0.428449259	5.35 E-08	0.090775697	7.88E-09
rs7531867	1	65880134	0.45169564	5.78E-08	0.210478443	1.15E-08
Rs2889195	1	65929318	0.442192407	1.40E-07	0.202793671	2.66E-08

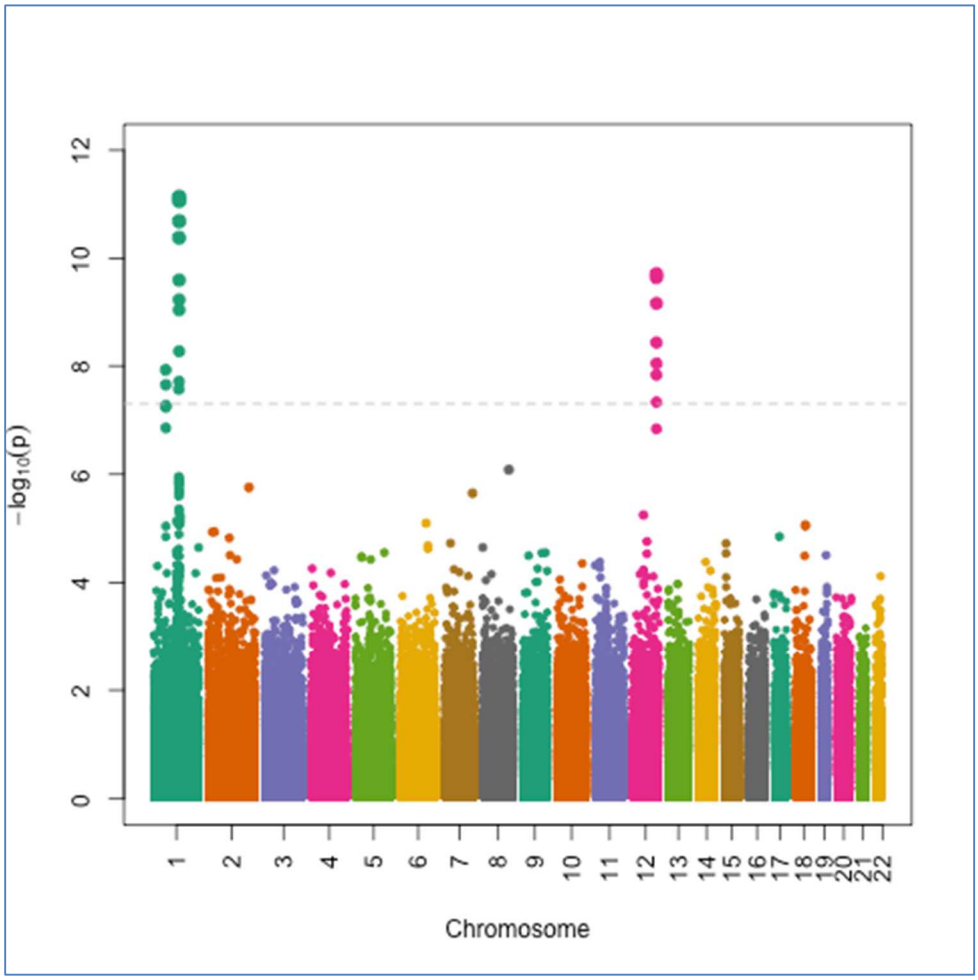
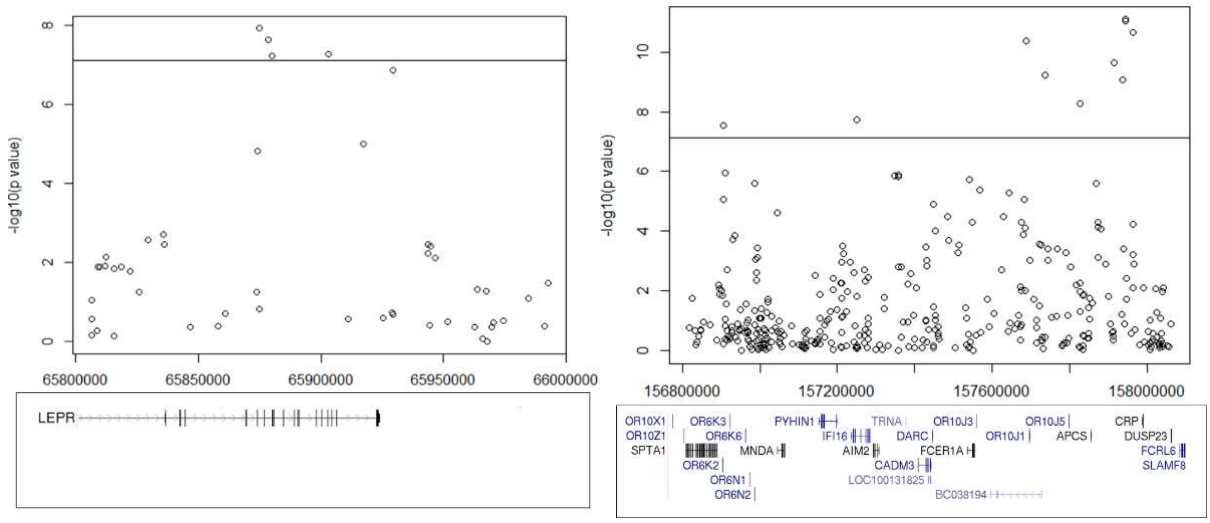


Figure 4.6. Manhhaton plot of MPQ test. The dash horizonal line indicates the genome-wide significance threshold that is based on the Bonferroni correction.



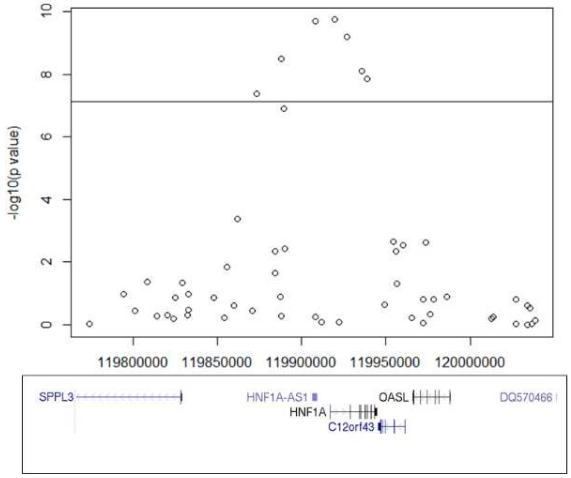


Figure 4.7. P values of SNPs inside 3 loci detected in MPQ test. Genes are shown at the bottom. Top left: chr1:65774845-66002986; Top right: chr1:156806658-158065173; Bottom right: chr12:119773345-120038696.

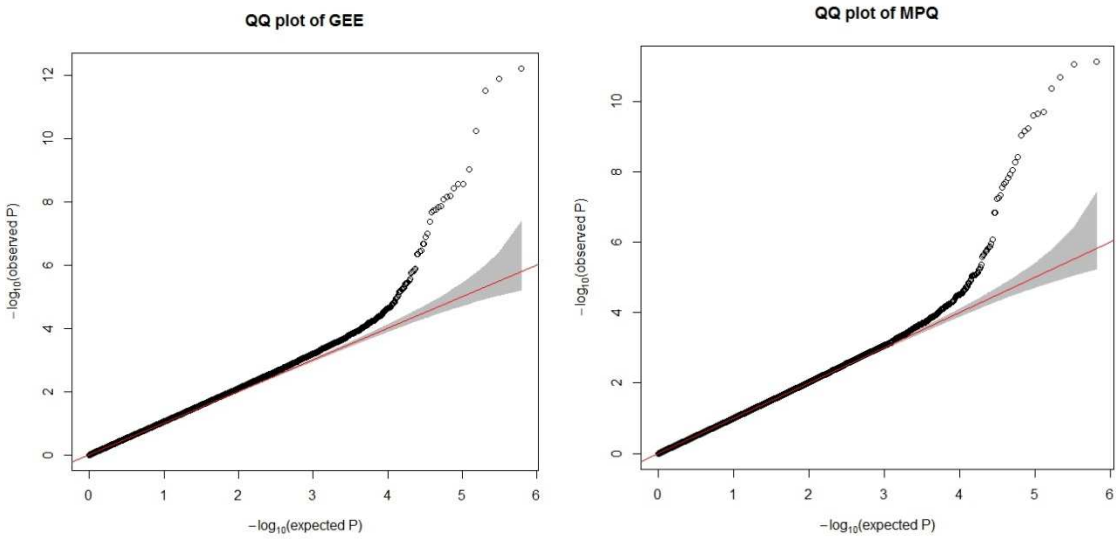


Figure 4.8. QQ plot of GEE test and MPQ test. Left: GEE; Right: MPQ.

Chapter 5

Discussion

The identification of genetic variants that are associated with multiple phenotypes can lead to a better understanding of the underpinnings of complex traits. For genetic association testing with multiple traits in samples that include related individuals, correlated genotypes and phenotypes must be appropriately accounted for to protect against spurious association. Careful considerations of these correlations can also improve power. In this thesis, we developed the multivariate phenotype quasi-likelihood (MPQ) for genetic association in related samples. The MPQ test is applicable to general combinations of family and population-based samples and can be used to test for genetic associations with general quantitative traits, including combinations of binary and continuous phenotypes. A nice feature of the MPQ test is that it is based on a retrospective analysis, therefore does not require correct specification of the distributions of the traits, which may not be known and are often assumed, such as normality for quantitative traits.

In simulation studies with both unrelated and related individuals, we compared the performance of the MPQ test to the previously proposed multivariate association test including the multivariate QLS test and GEE. We evaluated the methods under a variety of causal models with correlated structures between the genetic marker and traits based on different causal pathways. We demonstrated that MPQ and QLS have empirical type I error rates that are not significantly different from the nominal level, while GEE is not properly calibrated. In unrelated samples, the MPQ and QLS tests are identical and give very similar results to a score test based on a retrospective GLM, which is expected because when the log likelihood function for the GLM is in the exponential family, it is identical to the quasi-likelihood function (Wedderburn, 1974). However, if there is relatedness among individuals, the MPQ, QLS, and GLM score test are not the same, and with the GLM not being a valid association test in this setting.

In samples with related individuals, we demonstrated that the MPQ test provides an overall, and in many cases, substantial, improvement in power over the QLS test, while

retaining a computational simplicity that makes it useful in genome-wide association studies in arbitrary pedigrees. In simulation settings where multiple traits are not strongly correlated, the QLS generally has poor power, whereas the MPQ test has high power in this setting. We also demonstrate the performance in power of the QLS was highly dependent on the trait models and how the traits interacted, while the MPQ maintained consistently high power across models, suggesting that the MPQ test can be used in settings with complex interactions among traits.

The MPQ also offers several advantages over the QLS test. First, the MPQ allows for adjustment of covariates in the analysis. Second, the MPQ can also easily incorporate variance components and random effects to account for both genetic and non-genetic effects that are influencing the phenotypes. Finally, the MPQ can easily be extended to samples with unknown population and pedigree structure by using an empirical genetic relatedness matrix, calculated from genome-screen data, in the association analysis to account for unknown sample structure.

We applied the MPQ test to WHI-SHARe Hispanic cohort for the identification of pleiotropic effects for WBC count and CRP. The GWAS results of CRP with the WHI-SHARe Hispanic data have been published previously, and the study was performed under an additive genetic model with use of covariate-adjusted linear regression (Reiner, 2012). In this study, three loci were identified to have genome-wide significant associations with serum CRP level. In a GWAS of WBC count in 16388 African-American participants, DARC in 1q23 was shown to be associated with WBC count (Reiner, 2011). The MPQ test identified all of the significant associations that were previously detected in the univariate CRP GWAS. The MPQ test also has smaller p-values than the reported p-values for the univariate CRP study, indicating potential power gain with the multivariate MPQ. The MPQ test also identified a genome-wide significant association with the genetic variants in the DARC gene that was previously identified in African Americans. In addition, the MPQ identified two genome-wide significant SNPs that have not been previously reported and are not significant for the univariate models, thus demonstrating the utility of the multivariate MPQ test.

Despite the increase in degrees of freedom when testing multiple phenotypes with the MPQ, compared to a univariate test, we demonstrated in both simulation studies and real data applications that the joint analysis of multiple traits with MPQ can increase statistical power.

References

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

Chen MH, Liu X, Wei F, Larson MG, Fox CS, Vasan RS, Yang Q.A comparison of strategies for analyzing dichotomous outcomes in genome-wide association studies with general pedigrees. Genet Epidemiol. 2011;35(7):650-7.

Feng Z. (2014) A generalized Quasi-likelihood scoring approach for simultaneously testing the genetic association of multiple traits. Appl Stat. 63(3):483-498

Gibbons RD, Bock RD (1987). Trend in correlated proportions. Psychometricka 52, 112-124

Heyde, C. (1997) Quasi-likelihood and Its Application: a General Approach to Optimal Parameter Estimation. New York: Springer

Jakobsdottir J and McPeek S (2013). MASTOR: Mixed-Model Association Mapping of Quantitative Traits in Samples with Related individuals. Am J Human Genet. 92:652-666

Jia P, Zhao Z. Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. Hum Genet. 2013. 133(2):125-138

Karasik D, Cheung CL, Zhou Y, Cupples LA, Kiel DP, Demissie S. Genome-wide association of an integrated osteoporosis-related phenotype: is there evidence for pleiotropic genes? J Bone Miner Res. 2012;27(2):319-30.

Klei L, Luca D, Devlin B, and Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genetic Epidemiology, 2008; 32(1):9–19.

Klei L1, Luca D, Devlin B, Roeder K.Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet Epidemiol. 2008 Jan; 32(1):9-19.

Kung-Yee Liang and Scott Zeger. Longitudinal data analysis using generalized linear models. 1986; Biometrika 73 (1): 13–22

Lange C, Silverman EK, Xu X, et al. 2003. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostatistics 4:195-206

Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34, 816–834

Miall WE, Oldham PD. The hereditary factor in arterial blood-pressure. Br Med J. 1963 Jan 12; 1(5323):75–80.

Murabito JM, Rosenberg CL, Finger D, Kreger BE, Levy D, Splansky GL, Antman K, Hwang SJ. A genome-wide association study of breast and prostate cancer in the NHLBI's Framingham Heart Study. BMC Med Genet. 2007;8 Suppl 1:S6.

Ott J and Rabinowitz D, A principal-components approach based on heritability for combining phenotype information. *Human Heredity*, 1999. 49(2):106–111.

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA. Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol Ecol. 2012; 21(12):2991-3005.

Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine 1901; 2 (11): 559–572.

Pikula A, Beiser AS, Wang J, et al. Lipid and lipoprotein measurements and the risk of ischemic vascular events: Framingham Study. Neurology. 2015 Feb 3;84(5):472-9.

Reiner AP, Beleza S, Franceschini N, et al. Genome-wide Association and Population Genetic Analysis of C-Reactive Protein in African American and Hispanic American Women. Am J Hum Genet. 2012 September 7; 91(3): 502–512.

Reiner AP, Lettre G, Nalls MA, et al. Genome-Wide Association Study of White Blood Cell Count in 16,388 African Americans: the Continental Origins and Genetic Epidemiology Network (COGENT). PLoS Genet. 2011 June; 7(6): e1002108.

Sivakumaran S, Agakov F, Theodoratou E, et al. Abundant Pleiotropy in human complex

diseases and traits. Am J Human Genet. 2011, 89:607-618

Spielman RS, McGinnis RE, Ewens WJ. "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)". Am J Hum Genet. 1993; 52 (3): 506–16.

Stearns, F. W. One hundred years of pleiotropy: a retrospective. *Genetics* 186, 767–773 (2010)

Szymańska K, Hainaut P.TP53 and mutations in human cancer. Acta Biochim Pol. 2003;50(1):231-8

Thornton, T, and McPeek, M.S. (2010). ROADTRIPS: Casecontrol association testing with partially or completely unknown population and pedigree structure. Am. J. Hum. Genet. 86, 172–184.

van der Sijde MR, Ng A, Fu J. Systems genetics: From GWAS to disease pathways. Biochim Biophys Acta. 2014;1842(10):1903-1909.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five Years of GWAS Discovery. American Journal of Human Genetics 2012; 90:7-24.

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

Weil WB Jr. Juvenile diabetes mellitus. N Engl J Med. 1968 Apr 11;278(15):829-31.

Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetics association studies. J Prob Stat. 2012,

Zhu W and Zhang H. Why Do We Test Multiple Traits in Genetic Association Studies? J Korean Stat Soc . 2009 ; 38(1): 1–10