**이학박사학위논문**

**Statistical Method Development for   
Genetic Association Analyses of   
Dichotomous Phenotypes with Related Samples   
and its Application to Genetic Studies**

**종속 표본에 대한   
이분형 표현형의 유전체 연관성 분석 방법의 개발 및   
유전자 데이터에의 적용**

**2019년 2월**

**서울대학교 대학원**

**협동과정 생물정보학과**

**김 원 지**

**Statistical Method Development for   
Genetic Association Analyses of   
Dichotomous Phenotypes with Related Samples   
and its Application to Genetic Studies**

**by**

**Wonji Kim**

**A thesis   
submitted in fulfillment of the requirement   
for the degree of Doctor of Philosophy   
in   
Bioinformatics**

**Interdisciplinary Program in Bioinformatics**

**College of Natural Sciences**

**Seoul National University**

**Feb, 2019**

**Statistical Method Development for   
Genetic Association Analyses of   
Dichotomous Phenotypes with Related Samples   
and its Application to Genetic Studies**

**지도교수 원 성 호**

**이 논문을 이학박사 학위논문으로 제출함**

**2019년 2월**

**서울대학교 대학원**

**생물정보협동과정 생물정보학 전공**

**김 원 지**

**김원지의 이학박사 학위논문을 인준함**

**2018년 12월**

**위 원 장 박 태 성 (인)**

**부위원장 원 성 호 (인)**

**위 원 성 주 헌 (인)**

**위 원 유 연 주 (인)**

**위 원 이 우 주 (인)**

**Abstract**

**Statistical Method Development for   
Genetic Association Analyses of   
Dichotomous Phenotypes with Related Samples   
and its Application to Genetic Studies**

Wonji Kim

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

Recent improvements in sequencing technology have enabled the investigation of so-called “missing heritability”, and a large number of affected subjects have been sequenced in order to detect significant associations between human diseases and rare variants. However, the cost of genome sequencing is still high, and a statistically powerful strategy for selecting informative subjects would be useful. Moreover, numerous methods for estimating heritability have been proposed; however, unlike quantitative phenotypes, heritability estimation for dichotomous phenotypes is computationally and statistically complex, and the use of heritability is infrequent.

In this study, we propose a new statistical method for selecting cases and controls for sequencing studies based on disease family history. We assume that disease status is determined by unobserved liability score. Our method consists of two steps: first, the conditional means of liability are estimated given the individual’s disease status and those of their relatives with the liability threshold model, and second, the informative subjects are selected with the estimated conditional means. Our simulation studies showed that statistical power is substantially affected by the subject selection strategy chosen, and power is maximized when affected (unaffected) subjects with high (low) risks are selected as cases (controls). The proposed method was successfully applied to genome-wide association studies for type-2 diabetes, and our analysis results reveal the practical value of the proposed methods.

In addition, we developed a statistical method to estimate heritability of dichotomous phenotypes using a Liability Threshold Model in the context of ascertained family-based samples. The Liability Threshold Model assumes dichotomous phenotypes are determined by unobserved latent variables that are normally distributed, and this model can be applied to general pedigree data. The proposed methods were applied to simulated data and Korean type-2 diabetes family-based samples, and the accuracy of estimates provided by the experimental methods was compared with that of established methods.

**Key words**: Genome-wide association studies (GWAS), Family history of disease, Risk Prediction, Heritability, Liability threshold model, Ascertainment bias

**Student number**: 2015-30118

**Contents**

Contents

[Chapter 3 Selecting Cases and Controls for Genome-wide Association Studies Using Family Histories of Disease 9](#_Toc531786643)

[3.1 Introduction 9](#_Toc531786644)

# Chapter 3

**Selecting Cases and Controls for Genome-wide Association Studies Using Family Histories of Disease**

## 3.1 Introduction

Over the last several decades, DNA sequencing technologies have greatly improved, and the rate of decline in sequencing costs has even outpaced Moore’s law (MOORE 1998, Mardis 2008, Metzker 2010, Sboner, Mu et al. 2011). This progress has enabled well-powered investigations into the associations between human diseases and rare variants. Clues to the so-called “missing heritability” problem are also expected to emerge, as rare causal variants have been suggested as a possible cause (Maher 2008, Manolio, Collins et al. 2009). However, large-scale genetic association analyses often suffer from extreme multiple testing problems, and the cost of whole-genome sequencing is still expensive. Furthermore, the common disease-rare variant hypothesis (Pritchard 2001) assumes multiple rare disease susceptibility loci, suggesting that causal variants for each affected subject may be substantially different, and this genetic heterogeneity among affected subjects has also complicated genetic association analyses. Therefore, in spite of remarkable improvement in sequencing technology, development of efficient strategies for selecting informative subjects is still necessary, and various statistical methods have been investigated for use in genetic association studies.

Subjects with many affected relatives tend to contain more disease genotypes for heritable diseases, and it has been empirically shown that their ascertainment for genetic studies have often led to additional improvements in statistical power (Risch 2001, Antoniou and Easton 2003, Howson, Barratt et al. 2005, Li, Boehnke et al. 2006). In particular, the probability of being affected depends on both the number of affected/unaffected relatives and familial relationships. For instance, subjects with affected siblings have a greater chance of being affected than those with unaffected siblings, and the former rather than the latter are often selected for association analyses (Risch 2001, Antoniou and Easton 2003, Howson, Barratt et al. 2005, Li, Boehnke et al. 2006). Between subjects with three affected and one unaffected grandparent and those with a single affected parent, it is unclear which would be more efficient for genetic association studies. However, such complicated scenarios have rarely been considered due to the absence of appropriate statistical approaches, and many genetic association studies use only the number of affected first-degree relatives (Risch 2001, Antoniou and Easton 2003, Howson, Barratt et al. 2005, Li, Boehnke et al. 2006).

In this report, we propose a new statistical method for selecting informative subjects based on the disease status of their relatives. In our method, quantifying the how informative subjects are for association analyses requires knowing the prevalence and heritability of diseases *a priori*. In particular, prevalence is defined by the proportion of affected individuals in a population, and it is often available for many diseases. However, heritability for dichotomous phenotypes, which is defined by the proportion of the total phenotypic variance attributable to genetic components and estimated by familial correlation for quantitative phenotypes, can have different interpretations according to considered statistical models. For instance, heritability can be estimated from twin studies (Edwards 1969) or Falconer’s liability threshold model (Falconer 1965). The former estimates heritability through correlation of the disease status of monozygotic vs. dizygotic twins. The latter assumes that there are unobserved liability scores, and heritability is defined by correlation of liability scores, which can be understood as a correlation at the model scale (Stroup 2012), and some literature shows their asymptotic relationship (Lee, Wray et al. 2011). Heritability estimation at the observed data scale (Stroup 2012) is intuitively easier to understand, but its application to general family structures is not straightforward. Therefore, we consider heritability estimates from the liability threshold model in the remainder of this report.

Our model is based on the expectation of unobserved liability scores for subjects when the disease status of those subjects and their relatives are conditioned. The liability threshold model assumes that the disease status of each subject is affected if the unobserved liability score exceeds a threshold that is determined by prevalence; otherwise, the status is unaffected. It should be noted that this liability threshold model is equivalent to the probit model for independent samples (Bliss 1934). The unobserved liability scores are assumed to follow the normal distribution, and we calculate the conditional expectation with moment-based methods (Wilhelm and Manjunath 2013). The proposed method can utilize the disease status of any type of relative, and using extensive simulation studies, we show that the statistical power is maximized when subjects with high and low risk are selected as cases and controls, respectively. The proposed methods were applied to genome-wide association studies (GWAS) for type-2 diabetes (T2D) with data collected from the Korea Association REsource (KARE) project and Seoul National University Hospital in Korea (SNUH). The discovery of promising disease susceptibility loci illustrates the practical value of the proposed method.

## 3.2 Methods

### 3.2.1 Notations and the disease model

We assume that there are *n* independent subjects and that subject *i* has *ni* relatives (*i*=1, …, to *n*). We assume that the disease locus is biallelic, and denote normal and disease alleles by *d* and *D,* respectively. Their allele frequencies are assumed to be *pd* and *pD,* respectively. The genotypes are coded as the number of disease alleles, and genotype frequencies are assumed to follow Hardy-Weinberg equilibrium (HWE) in a population. We denote the genotypes of subject *i* and his/her relative *j* by *Gi* and *Gijr* respectively, and the genotype vectors are defined by

.

We consider the liability threshold model (Falconer 1965), and dichotomous phenotypes are determined by the unobserved continuous liability score. The liability scores of subject *i* and his/her relative *j* are denoted by *Li* and *Lijr,* respectively. The liability vector for relatives of subject *i* is denoted by

,

and that of both **L***i* and  is



We assume that liabilities are determined by summing the environmental effect, main genetic effect, polygenic effect, and random error. The environmental effects for subject *i* and his/her relatives are denoted by *Zi* and *Zijr*, and their vectors are defined by



Liability scores tend to be similar between family members, and we consider the simple additive polygenic effect model. We denote a *w*×*w* dimensional identity matrix by **I***w* and a *w* dimensional column vector, of which all elements are 0 and 1 by **0***w* and **1***w,* respectively. Then, if we let  and  be variances of polygenic effects and random residual effects, respectively, and let **Z***i* include the intercept, we can assume that



Here,  indicates the kinship coefficient matrix for both subject *i* and his/her relatives. We denote the kinship coefficient between subject *i* and his/her relative *j* by *πij* and that between two relatives *j* and *j'* by. Similarly, *di* and  denote the inbreeding coefficients for subject *i* and his/her relative *j*, respectively. The inbreeding coefficient, which ranges from 0 to 1, quantifies the departure from HWE and can be easily estimated using known pedigree by currently available R packages, e.g. *pedigreemm* (SARGOLZAEI and IWAISAKI 2005, Vazquez, Bates et al. 2010). Then,  and **Ψ***i* are defined by

 and .

Genomic relationships may have more information to better infer individual liability than the kinship coefficients. However, the genomic relationship matrix can be obtained only when the genotypes are known, which may not be the case in our study design.

The dichotomous phenotypes for subject *i* and his/her relative *j* are denoted by *Yi* and , respectively, and they are coded as 1 for cases and 0 for controls. In a liability threshold model, *Yi* and  are determined by *Li* and , respectively; if *Li* and  are above a certain threshold value *c*, *Yi* and  become 1, and otherwise they become 0. *c* can be determined from the prevalence of the diseases, and the phenotype vector for relatives of the subject *i* is denoted by

,

and that for the subject *i* and his/her relatives is denoted by



Several algorithms have been suggested to estimate *c* with prevalence, *q*, and heritability, *h*2, known *a priori*. For instance, if we denote the cumulative function of a standard normal distribution by Φ and there are no covariate effects other than the intercept, we can set *β*0 to be 0 without the loss of generality, and *c* can be obtained by the following equation:

.

If the environmental effect, *Z*, follows the normal distribution, and we denote its variance by, *c* can be obtained by

.

### 3.2.2 Selection of samples with extreme phenotypes

Subjects with extreme phenotypes lead to improvement of statistical power in genetic association studies (Risch and Zhang 1996, Nebert 2000, Perez-Gracia, Ruiz-Ilundain et al. 2002, Guey, Kravic et al. 2011, Li, Lewinger et al. 2011, Barnett, Lee et al. 2013), and association analyses have often been conducted with such subjects. At the sample selection stage, genotypes of subjects are not known, and we assume *β* = 0 in equation (1). In particular, environmental factors can affect the dichotomous phenotypes and if their effects are known, we can then define the adjusted extreme phenotypes for dichotomous phenotypes by the following conditional expectation (CE) of liability scores:

.

CEs were calculated with a moment-based method (Wilhelm and Manjunath 2013), and the detailed algorithm is provided in the Appendix. Once we calculated these for all subjects, *na* affected subjects with the largest CEs and *nu* unaffected subjects with the smallest CEs were selected for genetic association studies.

Computation of CEs assumes that *h*2 (heritability), *q* (prevalence), **Z**, and *β*0 are known. While *h*2, *q*, and **Z** are often available *a priori*, the regression coefficients of environment effects are usually estimated from logistic regression, and they cannot be used as estimates of *β*0in equation (1). For independent subjects, liability threshold models are equivalent to the generalized linear model with an inverse of a cumulative normal distribution as a link function, and if we assume that mean and variance for the cumulative normal distribution are 0 and 1.6, respectively, it is approximately equal to the logistic regression (Gelman and Hill 2006). Therefore, if we let

 and ,

regression coefficients from logistic regressions can be directly used as *β*0.

### 3.2.3 Statistical power when the family history of disease is controlled

The statistical power for genetic association analysis with a case-control study design can be calculated when the relatives’ phenotypes are conditioned. We consider the liability model in equation (1) and assume a major disease gene model. If we let *q* be the prevalence of the disease and we denote the genotype relative risks by

 and ,

under HWE, penetrances can be parameterized by

 ,

and .

The expected disease allele frequencies (DAFs) for the affected subject *i* and the unaffected subject *i'* are

and

.

If , both conditional probabilities can be simplified to

,

and otherwise,  can be numerically calculated. DAFs for case *i* and control *i*' can be obtained by

 and

.

Therefore, if we assume that there are *na* cases and *nu* controls and let

 and

,

the statistical power for a Cochran Armitage test (Cochran 1954, Armitage 1955) under the alternative hypothesis can be obtained from

.

If we denote the *α* quantile of the central chi-square distribution with a single degree of freedom by , the statistical power at significance level *α* becomes



## 3.3 Simulation study

### 3.3.1 The simulation model

We assume that there are *n* subjects with known phenotypes and that *na* cases and *nu* controls are selected among these for genotyping (*n* ≥ *na* + *nu*). We also assume that phenotypes for each subject’s relatives are available, and we consider three different scenarios: (1) phenotypes of two parents and four siblings are known; (2) phenotypes of four grandparents, two parents, and four siblings are known; and (3) phenotypes of two parents and four siblings are known for half of the subjects, and phenotypes of four grandparents, two parents, and four siblings are known for the other half. Pedigrees for scenarios 1 and 2 are provided in Figure 3.1. The *pD* was assumed to be 0.2, and genotype frequencies were obtained under HWE. Founders’ genotypes in each family were generated from B(2, *pD*), and the non-founders’ genotypes were obtained by randomly generated Mendelian transmissions. To generate phenotypes, we considered the disease model in equation (1). We assumed no environmental effect, and *β*0 was assumed to be 0. The polygenic effect and random errors for relatives of subject *i* were independently generated from the multivariate normal distribution with variances  and , respectively. The main genetic effect was obtained by the product of *β* and the number of disease alleles. If we let

and ,

 and *β* are obtained by the assumed *h*2 and . Here, *h*2 and  indicate the heritability and the relative proportion of variance explained by the disease genes. Once liabilities were generated, they were transformed into affected if larger than the threshold *c*, and otherwise were considered unaffected. The value of *c* was chosen to preserve the assumed prevalences of *q* = 0.1 or *q* = 0.2. For the evaluation of type-I errors and power, we assumed *ha*2 to be 0

**Figure 3.1. Family history of disease.** The person indicated by an arrow is a proband.



and 0.005, respectively, and *h*2 was assumed to be 0.2 and 0.4, respectively. If *ha*2 was set to 0, *β* became 0, which indicates the null hypothesis (no association between genetic variants and phenotypes). Empirical size and power estimates were calculated with 2,000 replicates at several significance levels. In each replicate, we assumed that *n* = 10,000, and both *na* and *nu* were assumed to be 500. Genetic association analyses were conducted under the assumption that genotypes were available only for *na* cases and *nu* controls.

We considered five different strategies for selecting cases and controls: (S1) cases and controls were randomly selected from affected and unaffected subjects, respectively; (S2) affected subjects with the highest CEs were selected as cases, and controls were randomly selected; (S3) affected subjects with the highest CEs and unaffected subjects with the lowest CEs were selected as cases and controls, respectively; (S4) cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls; and (S5) affected subjects with the lowest CEs and unaffected subjects with the highest CEs were selected as cases and controls, respectively. Moreover, for comparing the proposed method to a simple heuristic rule, we additionally considered another strategy (S6), where the largest (smallest) number of affected first-degree relatives was selected as cases (controls). And then, we compared empirical sizes and powers using logistic regression.

### 3.3.2 Evaluation of selection strategy with simulated data

We investigated the effect of the selection strategy with simulated data. Six strategies, S1 to S6, which we described in the Method section, were used for genetic association analyses and were performed with the logistic regression. For each strategy, we selected 500 cases and 500 controls from 10,000 individuals, and empirical type-I errors and power were evaluated for each scenario with 2,000 replicates. Quantile-quantile (QQ) plots (Figure 3.2) show that the nominal significance level was generally well preserved for scenario 1, and the empirical type-I error rates generally preserved the nominal significance level (Table 3.1). Figures 3.3-4 and Tables 3.2–3 show that the nominal significance levels were generally well preserved for scenarios 2 and 3 as well. Therefore, we can conclude that selection of cases and controls using CEs does not affect statistical validity.

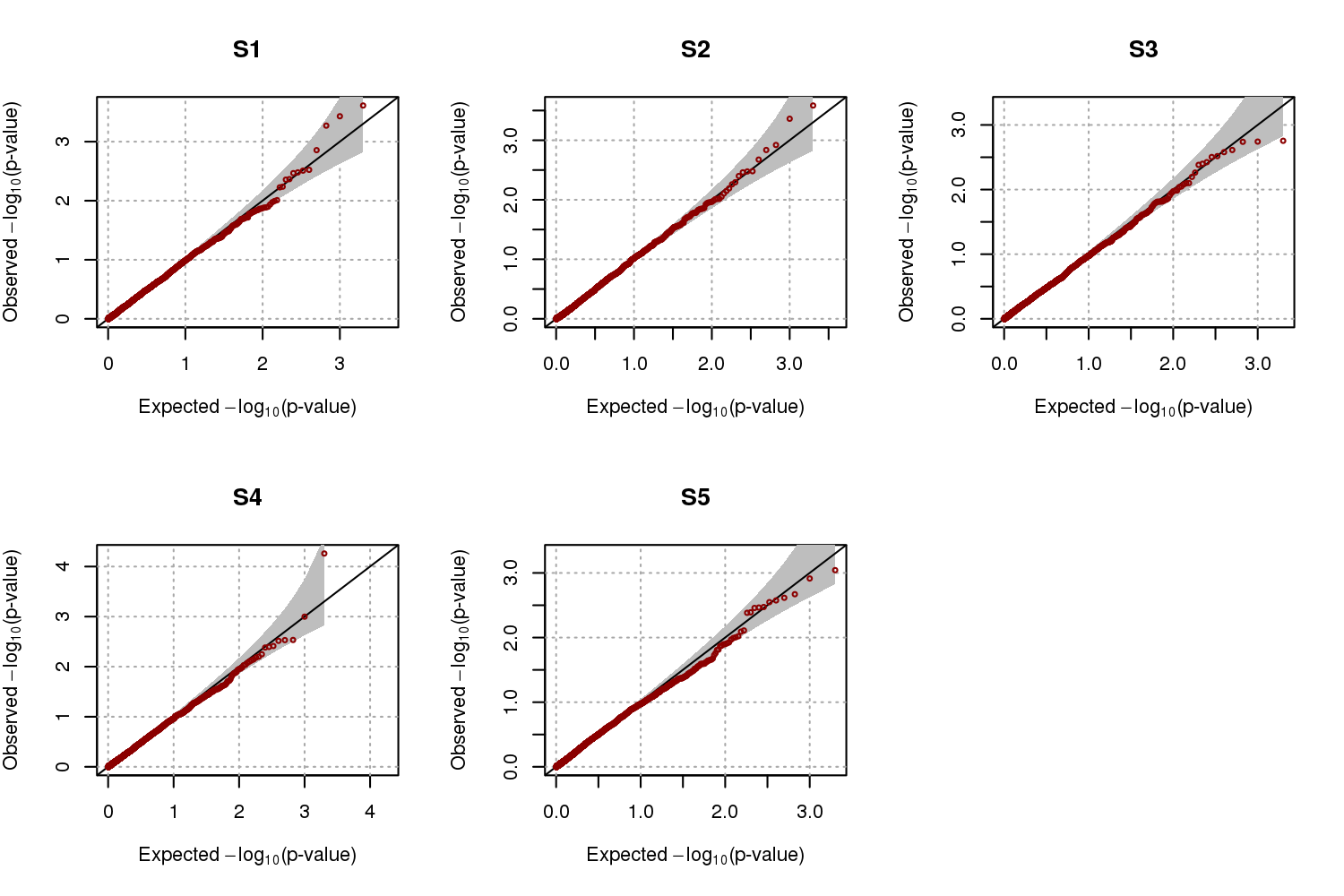
Empirical power levels were calculated at 0.005, 0.05, and 0.01 significance levels. We assumed that *ha*2 = 0.005, *h*2 = 0.2 or 0.4, and *q* = 0.1 or 0.2. Table 3.4 (scenario 1) shows that S3 was always the most efficient strategy among S1-S5, followed by S2 and S4. Interestingly, the statistical power estimates for S3 tended to be larger when the prevalence was larger and heritability was smaller, which indicates that the proposed method would be useful for common diseases. S5 always gave the highest rates of false-negative findings, as this strategy minimizes differences in DAFs between cases and controls. Table 3.5 (scenario 2) and Table 3.6 (scenario 3) showed very similar patterns to scenario 1. Therefore, we concluded that cases and controls ascertained with S3 leads to substantial improvement in power.

S6, the simple heuristic rule, showed an empirical power almost similar to that of S3 in scenario 1 (Table 3.4), i.e., S3 and S6 show no significant difference in performance when pedigrees are composed of only nuclear families with the same structure. However, since the proposed method considers not only the affected relatives, but also the unaffected relatives, S3 will be superior to S6 if many nuclear families of different structures are available. Moreover, S3 showed a better performance than S6 when pedigree structures were complex, as shown in Table 3.5 and Table 3.6, because S3 utilizes the disease status of all relatives, and not just first-degree ones. Therefore, as the degree of the known relatives increases, the proposed method gains strength because it uses all information, rather than being a simple heuristic rule.

**Table 3.1. Empirical type-I error estimates for scenario 1.** Scenario 1 was considered for family structures of subjects’ relatives. The empirical type-I errors were estimated with 2,000 replicates, and heritabilities were set to be 0.2 and 0.4.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *h*2 | *q* | Significance levels | S1a | S2b | S3c | S4d | S5e | S6f |
| 0.2 | 0.1 | 0.005 | 0.0055 | 0.0065 | 0.0040 | 0.0070 | 0.0050 | 0.0050 |
| 0.01 | 0.0070 | 0.0135 | 0.0090 | 0.0100 | 0.0105 | 0.0085 |
| 0.05 | 0.0515 | 0.0605 | 0.0510 | 0.0525 | 0.0555 | 0.0430 |
| 0.2 | 0.005 | 0.0020 | 0.0050 | 0.0040 | 0.0070 | 0.0070 | 0.0050 |
| 0.01 | 0.0050 | 0.0090 | 0.0100 | 0.0110 | 0.0115 | 0.0100 |
| 0.05 | 0.0395 | 0.0430 | 0.0550 | 0.0540 | 0.0520 | 0.0505 |
| 0.4 | 0.1 | 0.005 | 0.0045 | 0.0045 | 0.0050 | 0.0040 | 0.0060 | 0.0030 |
| 0.01 | 0.0090 | 0.0120 | 0.0115 | 0.0085 | 0.0145 | 0.0115 |
| 0.05 | 0.0440 | 0.0475 | 0.0450 | 0.0445 | 0.0495 | 0.0600 |
| 0.2 | 0.005 | 0.0050 | 0.0050 | 0.0045 | 0.0035 | 0.0070 | 0.0045 |
| 0.01 | 0.0110 | 0.0095 | 0.0085 | 0.0085 | 0.0105 | 0.0095 |
| 0.05 | 0.0555 | 0.0490 | 0.0460 | 0.0470 | 0.0510 | 0.0450 |
| aS1 : cases and controls were randomly selected from affected and unaffected subjects, respectively  bS2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected  cS3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)  dS4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls  eS5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)  fS6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls) | | | | | | | | |

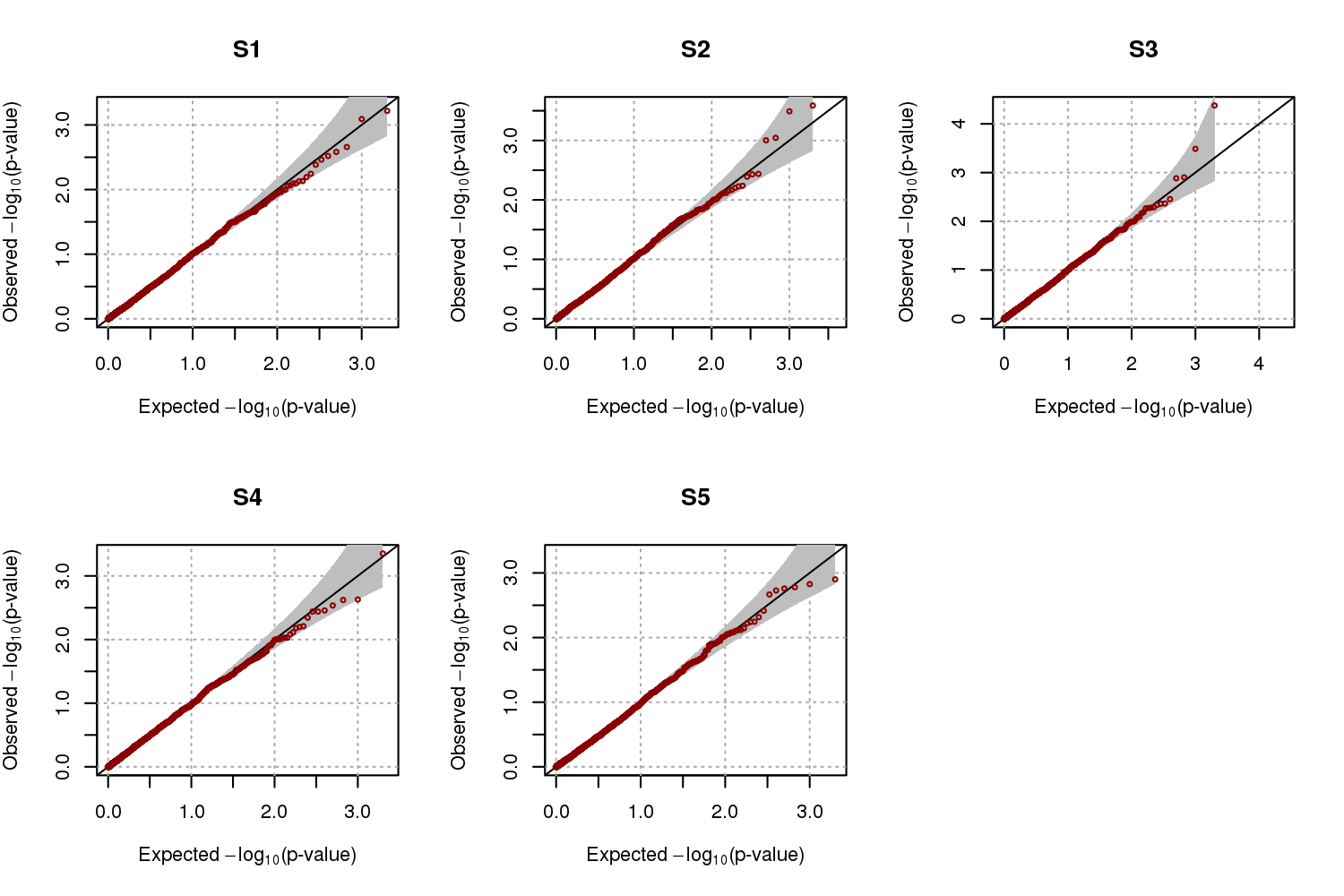
**Figure 3.2. Quantile-quantile (QQ) plots of simulated data for scenario 1.** We assume that and , and scenario 1 was assumed for relatives’ family structure. QQ plots were generated from 2,000 replicates

****

**Table 3.2. Empirical type-I error estimates for scenario 2.** Scenario 2 was considered for family structures of subjects’ relatives. The empirical type-I errors were estimated with 2,000 replicates, and heritabilities were set to be 0.2 and 0.4.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *h*2 | *q* | Significance levels | S1a | S2b | S3c | S4d | S5e | S6f |
| 0.2 | 0.1 | 0.005 | 0.0035 | 0.0035 | 0.0040 | 0.0040 | 0.0040 | 0.0045 |
| 0.01 | 0.0075 | 0.0095 | 0.0090 | 0.0095 | 0.0105 | 0.0095 |
| 0.05 | 0.0500 | 0.0560 | 0.0500 | 0.0500 | 0.0500 | 0.0420 |
| 0.2 | 0.005 | 0.0070 | 0.0030 | 0.0050 | 0.0065 | 0.0065 | 0.0045 |
| 0.01 | 0.0145 | 0.0095 | 0.0080 | 0.0095 | 0.0090 | 0.0110 |
| 0.05 | 0.0545 | 0.0415 | 0.0455 | 0.0460 | 0.0535 | 0.0540 |
| 0.4 | 0.1 | 0.005 | 0.0055 | 0.0090 | 0.0075 | 0.0045 | 0.0035 | 0.0055 |
| 0.01 | 0.0100 | 0.0155 | 0.0120 | 0.0090 | 0.0095 | 0.0100 |
| 0.05 | 0.0455 | 0.0555 | 0.0520 | 0.0420 | 0.0440 | 0.0375 |
| 0.2 | 0.005 | 0.0070 | 0.0050 | 0.0030 | 0.0035 | 0.0055 | 0.0065 |
| 0.01 | 0.0130 | 0.0100 | 0.0075 | 0.0065 | 0.0110 | 0.0110 |
| 0.05 | 0.0530 | 0.0570 | 0.0535 | 0.0500 | 0.0475 | 0.0550 |
| aS1 : cases and controls were randomly selected from affected and unaffected subjects, respectively  bS2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected  cS3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)  dS4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls  eS5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)  fS6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls) | | | | | | | | |

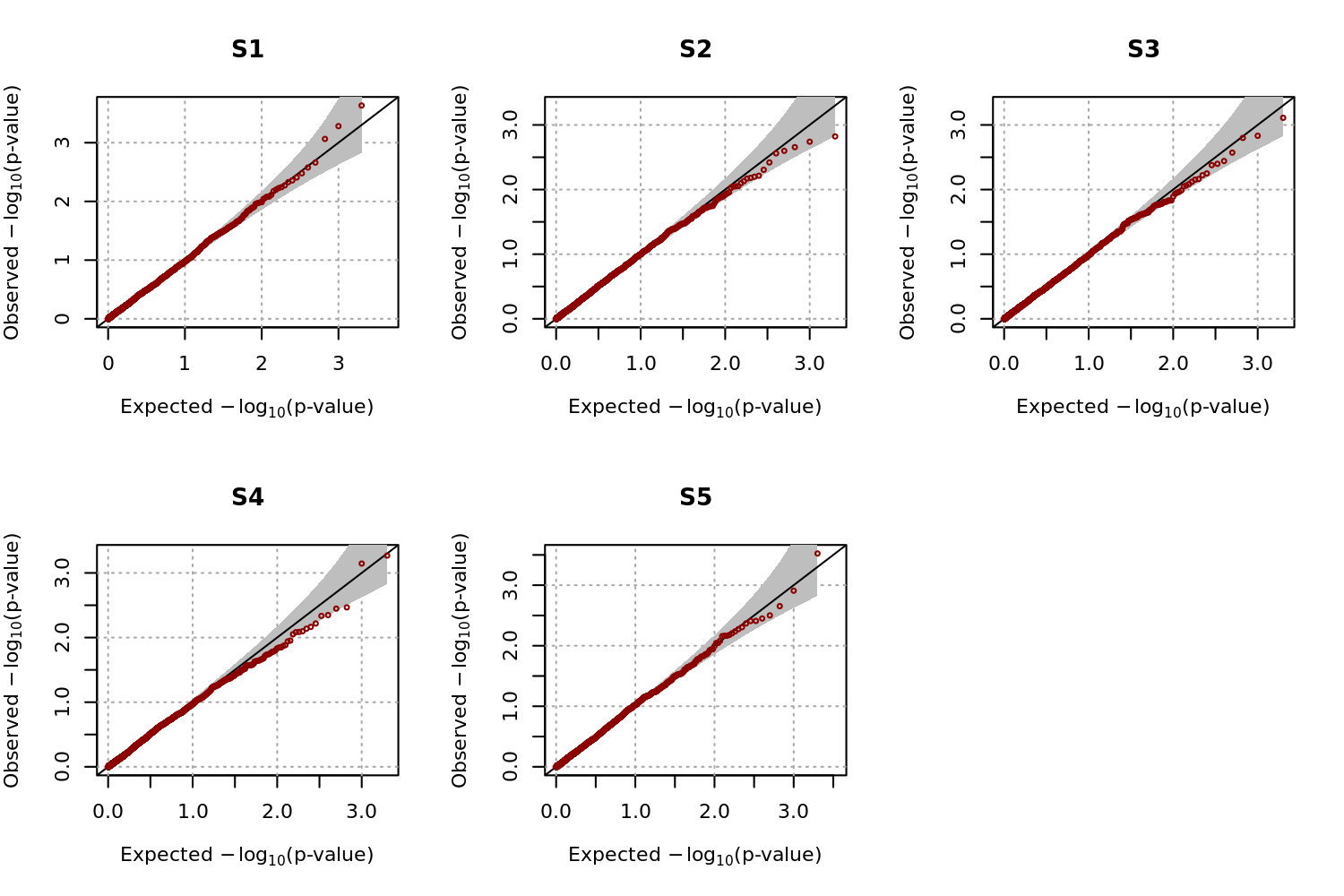
**Figure 3.3. Quantile-quantile (QQ) plots of simulated data for scenario 2.** We assume that and , and scenario 2 was assumed for relatives’ family structure. QQ plots were generated from 2,000 replicates

****

**Table 3.3. Empirical type-I error estimates for scenario 3.** Scenario 3 was considered for family structures of subjects’ relatives. The empirical type-I errors were estimated with 2,000 replicates, and heritabilities were set to be 0.2 and 0.4.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *h*2 | *q* | Significance levels | S1a | S2b | S3c | S4d | S5e | S6f |
| 0.2 | 0.1 | 0.005 | 0.0050 | 0.0045 | 0.0030 | 0.0025 | 0.0035 | 0.0045 |
| 0.01 | 0.0070 | 0.0090 | 0.0080 | 0.0085 | 0.0085 | 0.0095 |
| 0.05 | 0.0470 | 0.0450 | 0.0580 | 0.0525 | 0.0515 | 0.0520 |
| 0.2 | 0.005 | 0.0040 | 0.0055 | 0.0060 | 0.0070 | 0.0065 | 0.0060 |
| 0.01 | 0.0075 | 0.0090 | 0.0105 | 0.0120 | 0.0135 | 0.0130 |
| 0.05 | 0.0420 | 0.0440 | 0.0570 | 0.0570 | 0.0495 | 0.0650 |
| 0.4 | 0.1 | 0.005 | 0.0060 | 0.0075 | 0.0055 | 0.0025 | 0.0050 | 0.0055 |
| 0.01 | 0.0095 | 0.0135 | 0.0105 | 0.0095 | 0.0115 | 0.0130 |
| 0.05 | 0.0450 | 0.0560 | 0.0480 | 0.0500 | 0.0515 | 0.0540 |
| 0.2 | 0.005 | 0.0055 | 0.0040 | 0.0060 | 0.0040 | 0.0045 | 0.0045 |
| 0.01 | 0.0085 | 0.0075 | 0.0120 | 0.0080 | 0.0085 | 0.0100 |
| 0.05 | 0.0475 | 0.0450 | 0.0460 | 0.0480 | 0.0455 | 0.0490 |
| aS1 : cases and controls were randomly selected from affected and unaffected subjects, respectively  bS2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected  cS3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)  dS4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls  eS5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)  fS6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls) | | | | | | | | |

**Figure 3.4. Quantile-quantile (QQ) plots of simulated data for scenario 3.** We assume that and , and scenario 3 was assumed for relatives’ family structure. QQ plots were generated from 2,000 replicates

****

**Table 3.4. Empirical power estimates for scenario 1.**The empirical power levels were estimated with 2,000 replicates at different levels of significance. We assumed that *ha*2=0.005, *h*2 = 0.2 and 0.4, and *q* = 0.1 and 0.2.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *h*2 | *q* | Significance levels | S1a | S2b | **S3c** | S4d | S5e | **S6f** |
| 0.2 | 0.1 | 0.005 | 0.2675 | 0.4820 | **0.6635** | 0.4255 | 0.0030 | **0.6645** |
| 0.01 | 0.3505 | 0.5795 | **0.7450** | 0.5245 | 0.0085 | **0.7450** |
| 0.05 | 0.5880 | 0.8070 | **0.8980** | 0.7545 | 0.0520 | **0.8980** |
| 0.2 | 0.005 | 0.2210 | 0.5520 | **0.8220** | 0.4825 | 0.0095 | **0.8265** |
| 0.01 | 0.2840 | 0.6515 | **0.8815** | 0.5745 | 0.0195 | **0.8810** |
| 0.05 | 0.5260 | 0.8480 | **0.9645** | 0.7790 | 0.0930 | **0.9670** |
| 0.4 | 0.1 | 0.005 | 0.2700 | 0.4445 | **0.6090** | 0.4325 | 0.0085 | **0.6090** |
| 0.01 | 0.3525 | 0.5285 | **0.6925** | 0.5130 | 0.0155 | **0.6915** |
| 0.05 | 0.5950 | 0.7640 | **0.8670** | 0.7530 | 0.0675 | **0.8660** |
| 0.2 | 0.005 | 0.1825 | 0.4730 | **0.7010** | 0.4210 | 0.0055 | **0.6935** |
| 0.01 | 0.2425 | 0.5625 | **0.7825** | 0.5005 | 0.0135 | **0.7780** |
| 0.05 | 0.4725 | 0.7855 | **0.9215** | 0.7210 | 0.0530 | **0.9225** |
| aS1 : cases and controls were randomly selected from affected and unaffected subjects, respectively  bS2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected  cS3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)  dS4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls  eS5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)  fS6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls) | | | | | | | | |

**Table 3.5. Empirical power estimates for scenario 2.** The empirical power levels were estimated with 2,000 replicates at different levels of significance. We assumed that *ha*2=0.005, *h*2 = 0.2 and 0.4, and *q* = 0.1 and 0.2.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *h*2 | *q* | Significance levels | S1a | S2b | **S3c** | S4d | S5e | S6f |
| 0.2 | 0.1 | 0.005 | 0.2715 | 0.4960 | **0.7275** | 0.5165 | 0.0070 | 0.6730 |
| 0.01 | 0.3555 | 0.5855 | **0.7970** | 0.6160 | 0.0110 | 0.7565 |
| 0.05 | 0.6115 | 0.8010 | **0.9320** | 0.8240 | 0.0415 | 0.9030 |
| 0.2 | 0.005 | 0.1930 | 0.5940 | **0.9000** | 0.5485 | 0.0165 | 0.8115 |
| 0.01 | 0.2750 | 0.6840 | **0.9310** | 0.6530 | 0.0270 | 0.8685 |
| 0.05 | 0.5030 | 0.8595 | **0.9775** | 0.8415 | 0.0960 | 0.9565 |
| 0.4 | 0.1 | 0.005 | 0.2630 | 0.4355 | **0.6425** | 0.4625 | 0.0060 | 0.5850 |
| 0.01 | 0.3540 | 0.5285 | **0.7320** | 0.5585 | 0.0120 | 0.6795 |
| 0.05 | 0.5955 | 0.7495 | **0.8930** | 0.7875 | 0.0555 | 0.8720 |
| 0.2 | 0.005 | 0.1910 | 0.5080 | **0.7940** | 0.4870 | 0.0050 | 0.7185 |
| 0.01 | 0.2695 | 0.5975 | **0.8520** | 0.5800 | 0.0080 | 0.7855 |
| 0.05 | 0.4985 | 0.8030 | **0.9525** | 0.7885 | 0.0480 | 0.9185 |
| aS1 : cases and controls were randomly selected from affected and unaffected subjects, respectively  bS2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected  cS3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)  dS4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls  eS5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)  fS6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls) | | | | | | | | |

**Table 3.6. Empirical power estimates for scenario 3.** The empirical power levels were estimated with 2,000 replicates at different levels of significance. We assumed that *ha*2=0.005, *h*2 = 0.2 and 0.4, and *q* = 0.1 and 0.2.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *h*2 | *q* | Significance levels | S1a | S2b | **S3c** | S4d | S5e | S6f |
| 0.2 | 0.1 | 0.005 | 0.2700 | 0.4970 | **0.7475** | 0.5180 | 0.0045 | 0.6645 |
| 0.01 | 0.3490 | 0.5825 | **0.8065** | 0.6075 | 0.0095 | 0.7495 |
| 0.05 | 0.5980 | 0.7950 | **0.9245** | 0.8120 | 0.0405 | 0.9065 |
| 0.2 | 0.005 | 0.2135 | 0.5635 | **0.8860** | 0.5770 | 0.0185 | 0.8030 |
| 0.01 | 0.2850 | 0.6505 | **0.9215** | 0.6595 | 0.0340 | 0.8605 |
| 0.05 | 0.5380 | 0.8385 | **0.9825** | 0.8565 | 0.1130 | 0.9600 |
| 0.4 | 0.1 | 0.005 | 0.2615 | 0.4455 | **0.6375** | 0.4470 | 0.0090 | 0.5935 |
| 0.01 | 0.3485 | 0.5330 | **0.7205** | 0.5390 | 0.0185 | 0.6810 |
| 0.05 | 0.5855 | 0.7570 | **0.8795** | 0.7710 | 0.0655 | 0.8450 |
| 0.2 | 0.005 | 0.2130 | 0.4695 | **0.7860** | 0.5025 | 0.0090 | 0.7125 |
| 0.01 | 0.2890 | 0.5775 | **0.8475** | 0.6005 | 0.0175 | 0.7905 |
| 0.05 | 0.5020 | 0.7890 | **0.9515** | 0.7990 | 0.0570 | 0.9225 |
| aS1 : cases and controls were randomly selected from affected and unaffected subjects, respectively  bS2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected  cS3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)  dS4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls  eS5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)  fS6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls) | | | | | | | | |

### 3.3.3 Robustness of CE to choices of prevalence and heritability

The proposed selection strategy requires heritability and prevalence estimates, and the efficiency of the selection strategy can depend on the accuracy of these estimates. Therefore, we evaluated the sensitivity of the proposed method to misspecification of *h*2 and *q* values using simulated data. We considered the family structures in scenario 3, and the DAF in the population was assumed to be 0.2. Phenotypes for 10,000 subjects were generated with *ha*2 = 0.005, *h*2= 0.3, and *q* = 0.3. To evaluate the effect of misspecified values for (*h*2, *q*), these values were set to (0.1, 0.1), (0.2, 0.2), (0.4, 0.4), and (0.5, 0.5) for calculating CEs. Table 3.7 shows the relative ratio of power estimates for misspecified *h*2 and *q* compared to the results when *h*2 and *q* are correctly specified, with a value of 100 indicating that the power estimates are not affected. Results showed that the effect of misspecification of *h*2 and *q* seems to be almost negligible, at least for the considered simulation models.

Furthermore, ascertained cases and controls remain unchanged as long as the ranks of calculated CEs among cases (and controls) stay the same. We calculated the correlations between orders of true CEs and those with misspecified *h*2 and *q*. Figure 3.5 gives the contour plot of these correlations. It shows that correlations were always greater than 0.998, even when there were substantial differences between the true and misspecified *h*2 and *q*. Therefore, we can conclude that the rank of CEs remains largely the same, regardless of the values of *h*2 and *q* used.

**Table 3.7. Empirical relative power estimates for misspecified heritabilities and prevalences for scenario 3.** The empirical power levels were estimated with 2,000 replicates at different levels of significance and the ratios of the power estimates from misspecified (*h*2, *q*) to those from the correctly defined (*h*2, *q*) were calculated as percentage. We assumed that *ha*2=0.005 and (*h*2, *q*) = (0.3, 0.3) for generating phenotypes. Four misspecified pairs of (*h*2, *q*) were considered.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *h*2 | *q* | Significance levels | S1a | S2b | S3c | S4d | S5e |
| 0.1 | 0.1 | 0.005 | 102.899 | 100.705 | 99.888 | 100.657 | 88.235 |
| 0.01 | 103.586 | 99.774 | 99.946 | 99.841 | 92.857 |
| 0.05 | 100.106 | 98.425 | 100.154 | 100.540 | 100.000 |
| 0.2 | 0.2 | 0.005 | 104.348 | 98.325 | 100.503 | 101.221 | 97.059 |
| 0.01 | 102.110 | 98.417 | 100.270 | 101.351 | 98.214 |
| 0.05 | 98.301 | 98.308 | 99.897 | 101.439 | 97.222 |
| 0.4 | 0.1 | 0.005 | 106.087 | 97.884 | 100.447 | 101.972 | 91.176 |
| 0.01 | 106.118 | 97.513 | 100.486 | 101.510 | 91.071 |
| 0.05 | 96.603 | 99.650 | 100.410 | 98.741 | 103.333 |
| 0.5 | 0.2 | 0.005 | 95.072 | 101.146 | 100.280 | 102.723 | 88.235 |
| 0.01 | 99.367 | 99.925 | 100.054 | 103.021 | 94.643 |
| 0.05 | 102.866 | 99.242 | 100.513 | 100.540 | 104.444 |
| aS1 : cases and controls were randomly selected from affected and unaffected subjects, respectively  bS2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected  cS3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)  dS4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls  eS5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls) | | | | | | | |

**Figure 3.5. Contour plot for the correlation between orders of conditional expectations (CEs) calculated from true and misspecified** . Orders of CEs were obtained for the various choices of heritability and prevalence, and their correlations with true orders were calculated. Data were generated from and ‘’ is a point where correlation is exactly 1.



## 3.4 Application to genome-wide association of type-2 diabetes

### 3.4.1 The KARE cohort

The KARE cohort was collected to construct an indicator of disease with genetic influences in an attempt to predict the occurrence of various diseases. There are 8,842 participants consisting of 4,183 males and 4,659 females, and they were recruited from two Korean community cohorts, Ansung and Ansan, both in the Gyeonggi Province of South Korea. Participants are 40 to 69 years old. In total, 1,179 subjects were diagnosed as having T2D by a standard guideline (glucose at baseline ≥ 126 mg/dL, glucose 120 minutes after the insulin challenge ≥ 200 mg/dL, or HbA1c ≥ 6.5%). The disease status of their relatives was collected by a survey from all participants, and 1,037 subjects (125 cases and 912 controls) answered that they have affected relatives. In total, there were 1,230 affected relatives available.

The 8,842 subjects were genotyped for 352,228 SNPs with the Affymetrix Genome-Wide Human SNP Array 6.0. In our genome-wide association studies, we discarded SNPs for which the HWE p-values were less than 10-5, the genotype call rates were less than 95%, and the minor allele frequencies (MAF) were less than 0.05. We also eliminated subjects with gender inconsistencies, whose identity by state (IBS) was more than 0.8, or whose call rates were less than 95%. As a result, 310,515 SNPs for 8,842 subjects were utilized for GWAS.

### 3.4.2 The SNUH data

T2D patients were diagnosed by World Health Organization criteria from Seoul National University Hospital (SNUH), and 681 subjects with positive family history of diabetes in first-degree relatives were preferentially included. The disease status of their relatives was obtained based on the recall of the proband. However, family members were encouraged to perform a 75 g oral glucose tolerance test, and subjects positive for a glutamic acid decarboxylase autoantibody test were excluded. In total, the disease statuses of 7,825 relatives were available, among which 2,875 subjects had T2D.

T2D patients were genotyped with the Affymetrix Genome-Wide Human SNP Array 5.0, and 480,589 SNP genotypes were obtained. The same quality control conditions were applied as for the KARE samples, and 189,610 SNPs and two subjects were excluded. In total, 679 subjects with 290,979 SNP genotypes were used for the association analyses.

### 3.4.3 Association analyses using the pooled data

We used the proposed method to select cases and controls from KARE and SNUH samples for genetic association analyses of T2D. There were a total of 9,523 subjects (8,842 subjects from KARE and 681 subjects from SNUH). We excluded variants for which HWE p-values were less than 10-5, missing rates were greater than 5%, or MAFs were less than 0.05 and subjects whose call rates were less than 95% or IBS was more than 0.8. The remaining 9,521 subjects with 272,795 SNP genotypes were used for the analyses, and phenotypes of 7,804 relatives were available.

In the Korean population, about 9.9% of adults over 30 years of age were expected to have T2D in 2009 (Kim 2011), and the heritability of T2D has been reported to be approximately 26% (Poulsen, Kyvik et al. 1999). Therefore, we set the prevalence and heritability values at 0.099 and 0.26, respectively, and calculated CEs for the 9,521 subjects using the T2D status of their relatives. Based on these CEs, we selected 1,000 cases and 4,000 controls with S1 and S3. To adjust for population substructure, we calculated a genetic relationship matrix and applied the EIGENSTRAT approach (Price, Patterson et al. 2006). We obtained the top ten principal component (PC) scores with the largest eigenvalues, and they were included as covariates. We also included sex, age, and squared age as covariates.

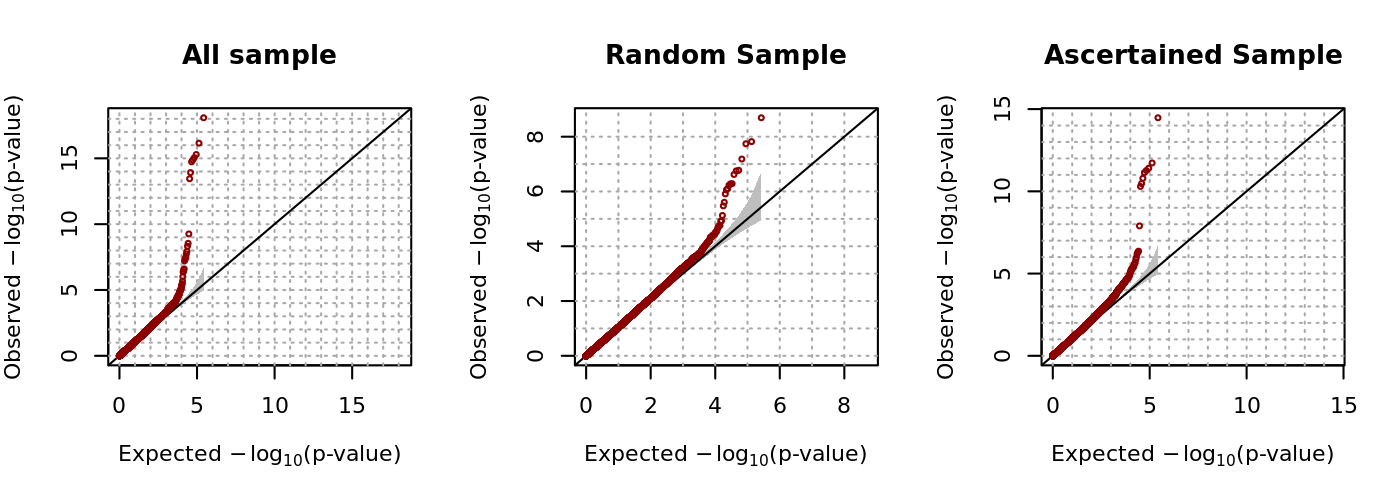
### 3.4.4 Results

We performed genome-wide association study for T2D using the pooled data to compare the performance between selection strategies which we considered in simulation study. The QQ-plots in Figure 3.6 show that GWAS using all subjects and using only the cases and controls ascertained with S1 and S3 preserve the nominal significance levels. Several studies showed that estimates from association analyses with cases and controls selected with family histories of diseases can be inflated (Bjørnland , Risch 2001, Antoniou and Easton 2003, Howson, Barratt et al. 2005, Li, Boehnke et al. 2006), and we conducted the other GWAS with permuted phenotypes. Figure 3.7 in supplementary file shows QQ-plots from GWAS with permuted phenotypes and we can conclude that statistical testing is robust against such problems. Figure 3.8 shows Manhattan plots for the analyses, with the genome-wide significance level adjusted by Bonferroni correction (1.872×10-7) indicated by dashed horizontal lines. The Manhattan plots reveal that the most significant results were obtained from GWAS using all subjects, followed by GWAS using cases and controls ascertained with S3. Table 3.8 shows results for SNPs that were significant in at least one of the GWAS analyses, and it has been reported in some researches that rs10946398, rs7754840, rs9465871, rs7747752, rs9348440, and rs10811661 are associated with T2D. Results showed that GWAS using cases and controls ascertained with S3 produced more significant SNPs than GWAS using cases and controls ascertained with S1. With the exception of rs10811661, p-values of all SNPs from the S3 GWAS were smaller than those from the S1 GWAS, and the genome-wide significance of SNPs from the S3 GWAS was much larger. Therefore, we can conclude that cases and controls ascertained with S3 leads to substantial improvement of power for GWAS.

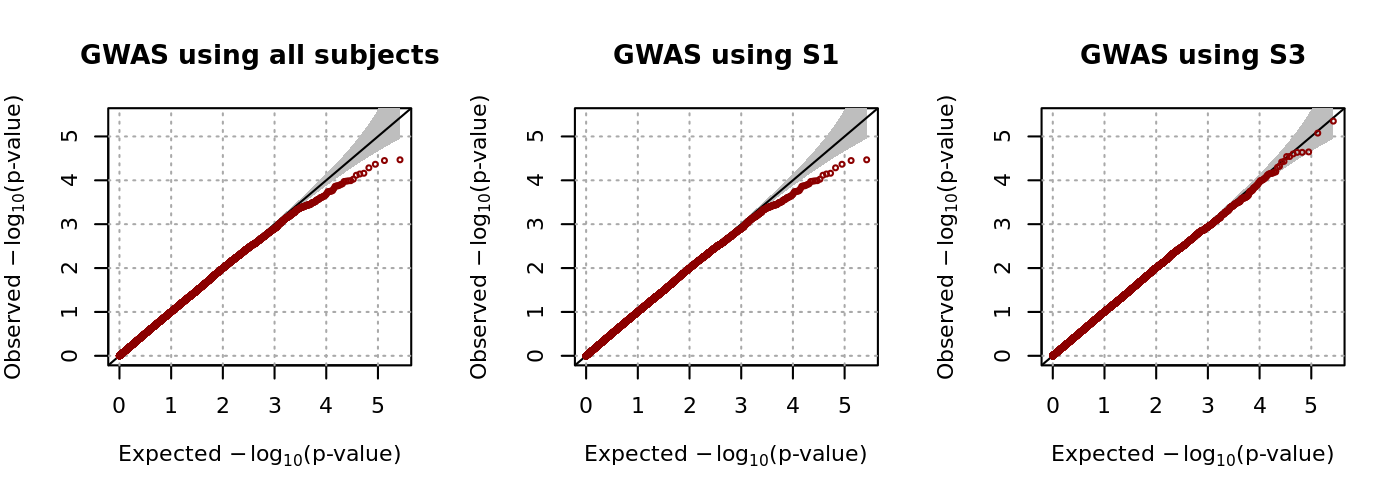
**Table 3.8. Results from GWAS.** The significance level adjusted by Bonferroni correction is 1.872×10-7 and significant SNPs are indicated in bold type.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SNP | CHR | POS | Gene | GWAS using all subjects | GWAS  using S1 | GWAS  using S3 |
| rs10946398 | 6 | 20661034 | CDKAL1 | **8.25×10-19** | **2.03×10-9** | **3.35×10-15** |
| rs7754840 | 6 | 20661250 | CDKAL1 | **7.03×10-17** | **1.82×10-8** | **1.88×10-12** |
| rs9460546 | 6 | 20663632 | CDKAL1 | **5.10×10-16** | **6.53×10-8** | **3.91×10-12** |
| rs9465871 | 6 | 20717255 | CDKAL1 | **8.91×10-16** | 2.40×10-7 | **1.61×10-11** |
| rs7747752 | 6 | 20725423 | CDKAL1 | **1.31×10-15** | **1.69×10-7** | **5.39×10-12** |
| rs7767391 | 6 | 20725240 | CDKAL1 | **1.84×10-15** | **1.78×10-7** | **7.21×10-12** |
| rs9348440 | 6 | 20641336 | CDKAL1 | **1.20×10-14** | 5.90×10-7 | **3.35×10-11** |
| rs2328549 | 6 | 20718240 | CDKAL1 | **3.53×10-14** | 2.48×10-6 | **5.02×10-11** |
| rs2328529 | 6 | 20631953 | CDKAL1 | **5.52×10-10** | 3.35×10-6 | 4.34×10-7 |
| rs10811661 | 9 | 22134094 | CDKN2B-AS1 | **2.84×10-9** | **1.51×10-8** | 1.04×10-6 |
| rs7741604 | 6 | 20731524 | CDKAL1 | **4.74×10-9** | 1.16×10-5 | 2.23×10-6 |
| rs1526959 | 12 | 79753790 | SYT1 | **1.16×10-8** | 3.00×10-3 | 2.89×10-6 |
| rs4291090 | 6 | 20570039 | CDKAL1 | **1.81×10-8** | 3.20×10-4 | 6.40×10-7 |
| rs2820001 | 6 | 20758943 | CDKAL1 | **3.23×10-8** | 9.19×10-5 | 2.05×10-5 |
| rs10946406 | 6 | 20758760 | CDKAL1 | **4.01×10-8** | 1.61×10-2 | 5.02×10-7 |
| rs2294809 | 6 | 20599888 | CDKAL1 | **4.52×10-8** | 4.90×10-4 | 2.41×10-6 |
| rs9366357 | 6 | 20599628 | CDKAL1 | **6.09×10-8** | 4.34×10-4 | 4.22×10-6 |
| rs12679402 | 8 | 41958980 | AP3M2 | 8.45×10-5 | 2.53×10-3 | **1.26×10-8** |

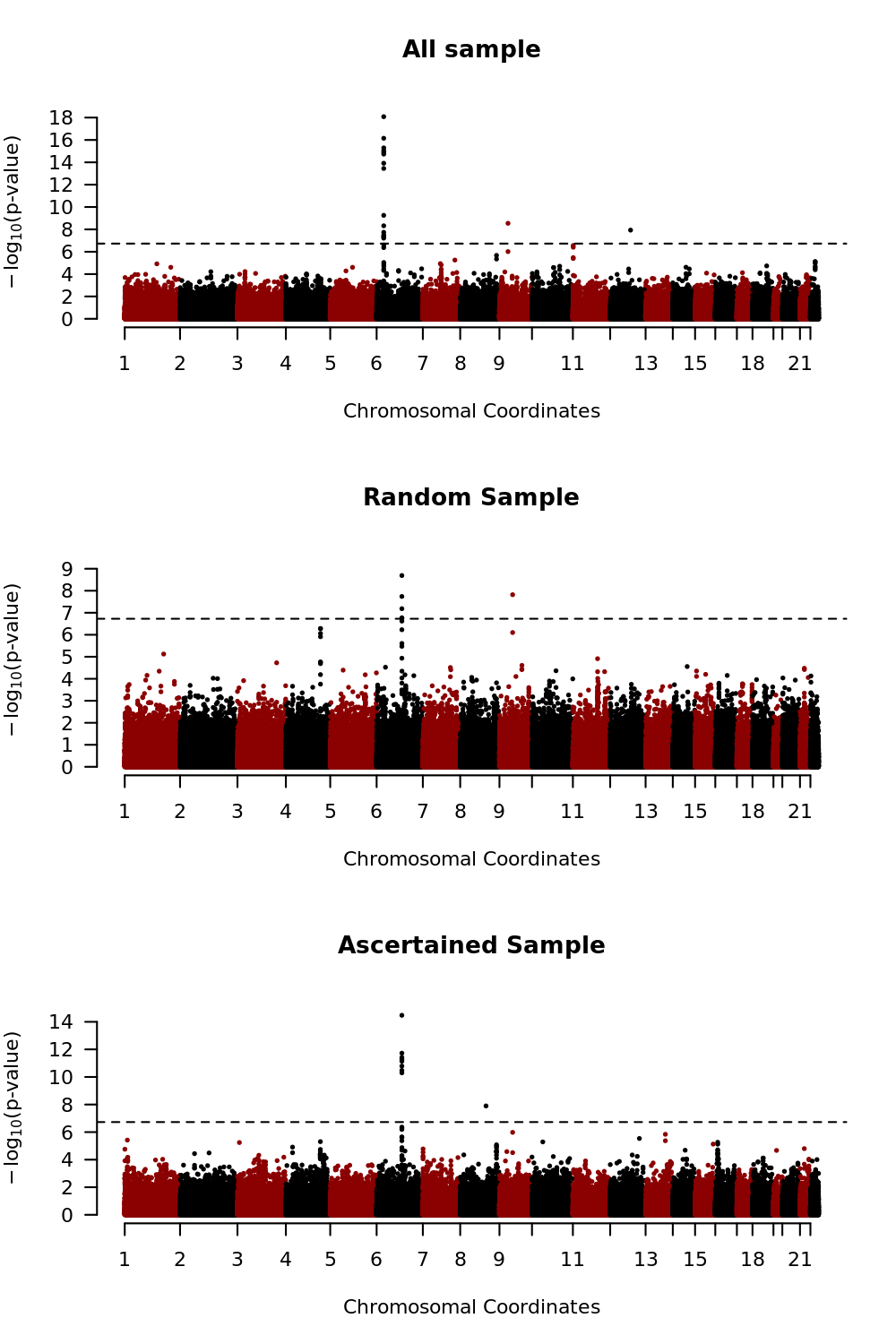
**Figure 3.6. Quantile-quantile plots for the results from genome-wide association study (GWAS) of type 2 diabetes.**

****

**Figure 3.7. QQ-plots for the GWAS with permuted phenotypes**



**Figure 3.8. Manhattan plots for the results from genome-wide association study (GWAS) of type 2 diabetes.**

****

## 3.5 Discussion

Many studies have reported that family history of a disease is related to statistical power (Risch 2001, Antoniou and Easton 2003, Howson, Barratt et al. 2005, Li, Boehnke et al. 2006). However, the effect of family history on genetic association analyses has not been carefully investigated, and its use for these analyses has been limited. For instance, subjects may be selected for genetic association analyses only if they have a certain number of affected relatives (Risch and Teng 1998). The effect of family history on genetic association analyses depends on the familial distance between relatives and the number of affected and unaffected relatives. In this report, we proposed a new statistical method for selecting the most informative cases and controls based on the family history of disease. The proposed method simultaneously takes into account both familial distance and number of relatives, and we show that selecting cases and controls using this method leads to a substantial improvement in statistical power. Our simulation results show that the improvement in statistical power tends to be larger for common and less heritable diseases. The proposed method was implemented using the R code, and it can accept various input file formats such as vcf, PLINK, and gen files. It can be downloaded free of cost from http://healthstat.snu.ac.kr/software/selSAMPLE.

Multiple studies have shown that subjects with extreme phenotypes lead to substantial improvement in statistical power (Li and Gail 2012, O’Reilly, Hoggart et al. 2012, Wang and Shete 2012, Schifano, Li et al. 2013, van der Sluis, Posthuma et al. 2013), and our proposed method can be considered as a statistical method to select such subjects with extreme phenotypes for dichotomous phenotypes. Association studies with extreme phenotypes were often utilized for continuous phenotypes (Risch and Zhang 1996, Nebert 2000, Perez-Gracia, Ruiz-Ilundain et al. 2002, Guey, Kravic et al. 2011, Li, Lewinger et al. 2011, Barnett, Lee et al. 2013), but it is not straightforward to define extreme phenotypes for dichotomous phenotypes. However, subjects with many affected relatives are expected to have higher liability scores, and thus, the presence of a higher number of affected relatives can be used to define extreme phenotypes. Alternatively, if there are continuous phenotypes correlated with the dichotomous phenotypes of interest, they can be utilized to define the extreme phenotypes. Extreme phenotypes can be defined in relation to those continuous phenotypes, and they can be utilized to select subjects. For instance, fasting glucose levels can be used to define extreme phenotypes for type-2 diabetes. These approaches can be used with existing software such as MTG2 (Lee and Van Der Werf 2016).

However, despite its flexibility, the proposed method has some limitations. First, our method involves the assumption that the liability scores follow a multivariate normal distribution; however, the estimated CEs may be biased if multivariate normality is violated (Benchek and Morris 2013). The generalized linear model can be understood as a latent variable model if its link function is an inverse function of some cumulative distribution (Bliss 1934). For instance, link functions for logistic and probit regressions are inverse functions of the cumulative logistic and standard normal distribution functions, respectively. Therefore, our liability threshold model can be considered as an extended probit model (Bliss 1934), and the distribution of unknown liability scores can be chosen by comparing several candidate link functions based on the Akaike information criteria (Bozdogan 1987). Second, there may be a recall bias for the family history of disease, and this bias could be substantial if accuracy is heterogeneous between cases and controls. Third, the proposed method requires that heritability and prevalence of the disease are known *a priori*. However, even if these values were unknown or incorrect, cases and controls selected with the proposed method would remain the same as long as the order of CEs among the affected and unaffected subjects was preserved. Alternatively, other approaches such as a generalized linear mixed model (GLMM) can be utilized to estimate the heritability and prevalence. For instance, GLMM can be applied with the family histories of diseases considered as responses. However, this method requires numerical integration, and its maximization becomes very complicated (McCulloch and Neuhaus 2001). Alternatively, we can consider the use of generalized estimating equations (Liang and Zeger 1986). However, family histories of diseases have a highly unbalanced structure, which often leads to slow or non-convergence of maximum likelihood estimations or to inflated statistical inferences (Fitzmaurice, Davidian et al. 2008). Therefore, further investigation is necessary. Fourth, estimates from a logistic regression would be unbiased if cases and controls were randomly selected from affected and unaffected subjects, respectively; however, if cases and controls are selected based on the family histories of the disease, it could lead to bias (Bjørnland). Fortunately, homogeneity tests between cases and controls are statistically valid as long as the estimates of odds ratio are carefully interpreted (Bjørnland).

Since the introduction of high throughput sequencing technology, substantial reductions in the cost for large-scale genetic association analyses have occurred, and many analyses have been launched to identify loci that show susceptibility. However, large-scale genetic analyses suffer from serious multiple-testing problems, and sequencing remains more expensive than phenotyping. Therefore, various statistical methods have been investigated to improve the power of testing. Our results reveal that additional statistical power can be achieved in association analyses with careful selection of cases and controls, and that the family history of disease is very useful for this purpose. Furthermore, the family history of disease is often obtained at relatively low costs, and therefore, the proposed method may be a useful strategy for improving the success of genome-wide association analyses.

## 3.6 Appendix

### 3.6.1 Calculation of the conditional expectation (CE)

Conditional expectation (CE) is derived with the moment-based approach with minor modifications (Wilhelm and Manjunath 2013). If we let *IA*(·) be an indicator function and define that

,

and , the CE for subject *i* is defined by

.

We use the moment-generating function (mgf) of the truncated multivariate normal distribution to calculate the conditional distribution. By definition, we can define the joint probability density function (pdf) of **L***i* by

, where .

The conditional pdf of **L***i* given  becomes



where . We can then find the mgf by



where. We let , and then the exponential term of mgf can be simplified to

,

and mgf becomes



We let *σijk* indicate the (*j*,*k*)th element of **Σ***i* and  indicate a marginal pdf for the *k*th element of **L***i* of the conditional pdf,, i.e.,

,

where subscript –*k* means that the *k*th element is removed from the corresponding vector.  will be derived in the next section. If we further denote



then the CE for subject *i* can be calculated by

.

### 3.6.2 Derivation of

The (*ni*+1)-dimensional liability vector, **L***i*, can be partitioned into (**L***i*)-*j* and *Lijr* for *j =* 1,…,*ni* or **L***ir* and *Li* for *j = ni*+1. For notational convenience, we only considered *j = ni*+1, which can be readily extended to the other subjects. The partitioned liability vector has the following distribution:

.

If we denote the lower and upper truncated points of **L***i* as **a***i* and **b***i* respectively, the truncated points for **L***i* are defined as

 and .

When **a***i* **< L***i* **< b***i*, the truncated normal distribution function is



By the property of multivariate normal distribution, the marginal pdf of *Li* at *Li* = *x* is given by

.

Because a conditional distribution of a normal distribution is also normally distributed, we know that **L***ir***|** *Li* **=** *x* is normally distributed with

 and .

Therefore, the multivariate marginal pdf of *Li* becomes



Here,  can be computed using statistical software, such as the function pmvnorm() in the R package mvtnorm.

# Chapter 4

**Heritability Estimation of Dichotomous Phenotypes Using a Liability Threshold Model on Ascertained Family-based Samples**

## 4.1 Introduction

Phenotypes are affected both by environmental factors and genes, and family members are expected to possess similar phenotypes due to their genetic similarity. Heritability was defined to quantify phenotypic similarity attributable to heritable components, and this concept has been widely used to understand the genetic architecture of phenotypes [Visscher, et al. 2008]. For example, heritability can be used to compare the importance of genetic components among different phenotypes. Additionally, if large-scale genetic data are available, genetic correlation matrices can be estimated [Fedko, et al. 2015]. These data can then be incorporated into a linear mixed model to provide SNP heritability estimation. SNP heritability provides information regarding the relative proportion of variance attributable to the genotyped SNPs, and this technique can be used to identify the degree of missing heritability.

Estimation of broad-sense heritability requires the study of bilinear relatives such as sibling or monozygotic twins, and in practice, narrow-sense heritability has often been utilized. Narrow-sense heritability is defined as the proportion of the total phenotypic variation explained by additive genetic effects [Visscher, et al. 2008]. Various methods have been developed for estimating the heritability of continuous traits. For example, restricted maximum likelihood methods based on the linear mixed model (LMM) [Vattikuti, et al. 2012; Yang, et al. 2010; Yang, et al. 2011] or polygenic score methods [Dudbridge 2013] can be used for estimating the heritability of continuous traits. For dichotomous traits, generalized linear mixed models or Liability Threshold Models have been often utilized [Burton, et al. 1999; Papachristou, et al. 2011]. The Liability Threshold Model assumes there are unobserved continuous liability scores, and subjects are affected if they exceed a certain threshold [Dempster and Lerner 1950; Hoeschele and Tier 1995; Lee, et al. 2011; Van Vleck 1972].

In this study, we focus on heritability estimation of dichotomous phenotypes. There are multiple factors which can bias variance estimation of dichotomous traits. In particular, family-based samples are typically analyzed using probands. The term proband refers to instances when family members are brought into a study as a result of other family members already enrolled in the study. Multiple reports indicate that proband analysis can produce substantial bias in variance estimates [Park, et al. 2015; Sawyer 1990; Yang, et al. 2011]. For example, if phenotypes are rare and families are randomly selected, the number of affected individuals is often very small. Therefore families are ascertained through the use of affected probands. In such instances, the majority of the relatives may be unaffected unless the size of the family is very large, and negative correlation can be observed because probands are affected while their relatives are unaffected. Several approaches have been proposed to adjust for such bias. GCTA adjusts estimated heritabilities by assuming that the level of ascertainment bias is same among individuals [Yang, et al. 2011]; however, families are ascertained with probands and the effect of ascertainment bias is heterogeneous according to familial relationship [Park, et al. 2015]. For example, ascertainment bias for grandparents of the proband is expected to be approximately half that of the parents.

Here, we developed a new method to estimate heritability based on the Liability Threshold Model for binary traits (LTMH) which can be applied to the extended pedigree structure. Using the Expectation-Maximization (EM) algorithm, the proposed method jointly estimates maximum likelihood estimators (MLE) for heritability and coefficients of covariates [Dempster, et al. 1977]. Furthermore, the proposed method maximizes the conditional likelihood of disease statuses of probands via a conditional EM (CEM) algorithm [Jebara and Pentland 1999], and ascertainment bias can be adjusted. We also developed a conditional expected score test (CEST) to determine if heritability is equal to zero. Extensive simulation studies demonstrated that heritability estimates obtained from the proposed methods are generally unbiased even for the ascertained family-based samples. Estimates from GCTA are unbiased for randomly selected families, but the bias turns out to be substantial for ascertained families. Also we found that the CEST for heritability was statistically conservative, but it could achieve reasonable statistical power estimates. Finally, we used the proposed method to estimate the heritability of type-2 diabetes (T2D) using ascertained family-based samples from Korean families, and those estimates confirmed the practical value of our proposed methods.

**Bibliography**

Abney, M., et al. (2000). "Estimation of variance components of quantitative traits in inbred populations." The American Journal of Human Genetics **66**(2): 629-650.

Antoniou, A. C. and D. F. Easton (2003). "Polygenic inheritance of breast cancer: implications for design of association studies." Genetic epidemiology **25**(3): 190-202.

Antoniou, A. C. and D. F. Easton (2003). "Polygenic inheritance of breast cancer: Implications for design of association studies." Genet Epidemiol **25**(3): 190-202.

Armitage, P. (1955). "Tests for linear trends in proportions and frequencies." Biometrics **11**(3): 375-386.

Atkinson, K. E. (2008). An introduction to numerical analysis, John Wiley & Sons.

Barnett, I. J., et al. (2013). "Detecting rare variant effects using extreme phenotype sampling in sequencing association studies." Genetic epidemiology **37**(2): 142-151.

Benchek, P. H. and N. J. Morris (2013). "How meaningful are heritability estimates of liability?" Human genetics **132**(12): 1351-1360.

Bertsekas, D. P. (2014). Constrained optimization and Lagrange multiplier methods, Academic press.

Bjørnland, T. "Statistical Methods for Genetic Association Studies under the Extreme Phenotype Sampling Design."

Bliss, C. I. (1934). "The method of probits." Science **79**(2037): 38-39.

Bozdogan, H. (1987). "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions." Psychometrika **52**(3): 345-370.

Cochran, W. G. (1954). "Some Methods for Strengthening the Common &#x3c7;<sup>2</sup> Tests." Biometrics **10**(4): 417-451.

Dempster, A. P., et al. (1977). "Maximum likelihood from incomplete data via the EM algorithm." Journal of the royal statistical society. Series B (methodological): 1-38.

Edwards, J. (1969). "Familial predisposition in man." British Medical Bulletin **25**(1): 58-64.

Falconer, D. S. (1965). "The inheritance of liability to certain diseases, estimated from the incidence among relatives." Annals of Human Genetics **29**(1): 51-76.

Finkelstein, D. M., et al. (2010). "A score test for association of a longitudinal marker and an event with missing data." Biometrics **66**(3): 726-732.

Fisher, R. A. (1919). "XV.—The correlation between relatives on the supposition of Mendelian inheritance." Earth and Environmental Science Transactions of the Royal Society of Edinburgh **52**(2): 399-433.

Fisher, R. A. (1925). Theory of statistical estimation. Mathematical Proceedings of the Cambridge Philosophical Society, Cambridge University Press.

Fitzmaurice, G., et al. (2008). Longitudinal data analysis, CRC Press.

Gelman, A. and J. Hill (2006). Data analysis using regression and multilevel/hierarchical models, Cambridge University Press.

Guey, L. T., et al. (2011). "Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants." Genetic epidemiology **35**(4): 236-246.

Howson, J. M., et al. (2005). "Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci." Genet Epidemiol **29**(1): 51-67.

Howson, J. M., et al. (2005). "Comparison of population‐and family‐based methods for genetic association analysis in the presence of interacting loci." Genetic epidemiology **29**(1): 51-67.

Jacquard, A. (2012). The genetic structure of populations, Springer Science & Business Media.

Kim, D. J. (2011). "The epidemiology of diabetes in Korea." Diabetes & metabolism journal **35**(4): 303-308.

Kuhn, H. W. and A. W. Tucker (2014). Nonlinear programming. Traces and emergence of nonlinear programming, Springer**:** 247-258.

Lee, S. H. and J. Van Der Werf (2016). "MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information." Bioinformatics **32**(9): 1420-1422.

Lee, S. H., et al. (2011). "Estimating missing heritability for disease from genome-wide association studies." The American Journal of Human Genetics **88**(3): 294-305.

Li, D., et al. (2011). "Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies." Genetic epidemiology **35**(8): 790-799.

Li, H. and M. H. Gail (2012). "Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies." Human heredity **73**(3): 159-173.

Li, M., et al. (2006). "Efficient study designs for test of genetic association using sibship data and unrelated cases and controls." Am J Hum Genet **78**(5): 778-792.

Li, M., et al. (2006). "Efficient study designs for test of genetic association using sibship data and unrelated cases and controls." The American Journal of Human Genetics **78**(5): 778-792.

Liang, K.-Y. and S. L. Zeger (1986). "Longitudinal data analysis using generalized linear models." Biometrika **73**(1): 13-22.

Maher, B. (2008). "The case of the missing heritability." Nature **456**(7218): 18-21.

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

Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends in genetics **24**(3): 133-141.

McCulloch, C. E. and J. M. Neuhaus (2001). Generalized linear mixed models, Wiley Online Library.

Metzker, M. L. (2010). "Sequencing technologies—the next generation." Nature reviews genetics **11**(1): 31-46.

MOORE, G. E. (1998). "Cramming More Components onto Integrated Circuits." PROCEEDINGS OF THE IEEE **86**(1).

Nebert, D. W. (2000). "Extreme discordant phenotype methodology: an intuitive approach to clinical pharmacogenetics." European journal of pharmacology **410**(2): 107-120.

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

Perez-Gracia, J. L., et al. (2002). "The role of extreme phenotype selection studies in the identification of clinically relevant genotypes in cancer research." CANCER-PHILADELPHIA- **95**(7): 1605-1610.

Poulsen, P., et al. (1999). "Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study." Diabetologia **42**(2): 139-145.

Price, A. L., et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics **38**(8): 904.

Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex diseases?" The American Journal of Human Genetics **69**(1): 124-137.

Risch, N. (2001). "Implications of multilocus inheritance for gene-disease association studies." Theor Popul Biol **60**(3): 215-220.

Risch, N. (2001). "Implications of multilocus inheritance for gene–disease association studies." Theoretical population biology **60**(3): 215-220.

Risch, N. and J. Teng (1998). "The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling." Genome Research **8**(12): 1273-1288.

Risch, N. J. and H. Zhang (1996). "Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations." American journal of human genetics **58**(4): 836.

SARGOLZAEI, M. and H. IWAISAKI (2005). "Comparison of four direct algorithms for computing inbreeding coefficients." Animal Science Journal **76**(5): 401-406.

Sboner, A., et al. (2011). "The real cost of sequencing: higher than you think!" Genome biology **12**(8): 1.

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

Stroup, W. W. (2012). Generalized linear mixed models: modern concepts, methods and applications, CRC press.

van der Sluis, S., et al. (2013). "TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies." PLoS Genet **9**(1): e1003235.

Vazquez, A., et al. (2010). "Technical note: an R package for fitting generalized linear mixed models in animal breeding." Journal of animal science **88**(2): 497-504.

Wang, J. and S. Shete (2012). "Analysis of secondary phenotype involving the interactive effect of the secondary phenotype and genetic variants on the primary disease." Annals of Human Genetics **76**(6): 484-499.

Wilhelm, S. and B. Manjunath (2013). "tmvtnorm: Truncated multivariate normal and student t distribution." R package version: 1.4-8.