# Selecting Cases and Controls for DNA Sequencing Studies Using Family Histories of Disease

Wonji Kim[1], Dandi Qiao[2], Michael H. Cho[2,3], Soo Heon Kwak[4], Kyoung Soo Park[4],

Edwin K Silverman[2,3], Pak Sham[5,6,7], Sungho Won[1,2,8,9,10*]

[1]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, Korea

[2]Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA 02115, USA

[3]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA.

[4]Department of Internal Medicine, Seoul National University, Seoul, Korea

[5]Department of Psychiatry, University of Hong Kong, Hong Kong SAR, China

[6]Genome Research Centre, University of Hong Kong, Hong Kong SAR, China

[7]State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong SAR, China

[8]Department of Public Health Science, Seoul National University, Seoul, Korea

[9]Institute of Health and Environment, Seoul National University, Seoul, Korea

[10]National Cancer Center, Goyang,Gyeonggi, Korea

[*]Corresponding Author:

Sungho Won, Department of Public Health Science, Seoul National University
1 Kwanak-ro Kwanak-gu Seoul 151-742 Korea
(Email) won1@snu.ac.kr, (Tel) +82-2-880-2714, (Fax) +82-303-0942-2714

## ABSTRACT

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. Therefore, in this report, 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.

## KEY WORDS

Family history of disease, risk prediction, liability

## RUNNING TITLES

Efficient strategy for subject selection

## 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 [1-4]. 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 [5, 6]. However, large-scale genetic association analyses often suffer from extreme multiple testing problems, and the cost of whole-genome sequencing is still quite expensive. Furthermore, the common disease-rare variant hypothesis [7] assumes multiple rare disease susceptibility loci, suggesting that causal variants for each affected subject may be quite 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 [8-11]. 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 [8-11]. 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 [8-11].

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 [12] or Falconer's liability threshold model [13]. 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 [14], and some literature shows their asymptotic relationship [15]. Heritability estimation at the observed data scale [14] 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 [16]. The unobserved liability scores are assumed to follow the normal distribution, and we calculate the conditional expectation with moment-based methods [17]. 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.

## MATERIALS AND METHODS

### Notations and disease model

We assume that there are $n$ independent subjects and that subject $i$ has $n_i$ 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 $p_d$ and $p_D$, 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 $G_i$ and $G_{ij}{}^r$ respectively, and the genotype vectors are defined by

$$\mathbf{G}_i^r = \begin{pmatrix} G_{i1}^r \\ \vdots \\ G_{in_i}^r \end{pmatrix} \text{ and } \mathbf{G}_i = \begin{pmatrix} \mathbf{G}_i^r \\ G_i \end{pmatrix}.$$

We consider the liability threshold model [18], 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 $L_i$ and $L_{ij}{}^r$, respectively. The liability vector for relatives of subject $i$ is denoted by

$$\mathbf{L}_i^r = \begin{pmatrix} L_{i1}^r \\ \vdots \\ L_{in_i}^r \end{pmatrix},$$

and that of both $\mathbf{L}_i$ and $\mathbf{L}_i^r$ is

$$\mathbf{L}_i = \begin{pmatrix} \mathbf{L}_i^r \\ L_i \end{pmatrix}.$$

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 $Z_i$ and $Z_{ij}{}^r$, and their vectors are defined by

$$\mathbf{Z}_i^r = \begin{pmatrix} Z_{i1}^r \\ \vdots \\ Z_{in_i}^r \end{pmatrix} \text{ and } \mathbf{Z}_i = \begin{pmatrix} \mathbf{Z}_i^r \\ Z_i \end{pmatrix}.$$

Liability scores tend to be similar between family members, and we consider the simple additive polygenic effect model. We denote a $w \times w$ dimensional identity matrix by $\mathbf{I}_w$ and a $w$ dimensional column vector, of which all elements are 0 and 1 by $\mathbf{0}_w$ and $\mathbf{1}_w$, respectively. Then, if we let $\sigma_g^2$ and $\sigma_e^2$ be variances of polygenic effects and random effects, respectively, and let $\mathbf{Z}_i$ include the intercept, we can assume that

$$\mathbf{L}_i = \mathbf{Z}_i\beta_0 + \mathbf{G}_i\beta + \mathbf{P}_i + \mathbf{E}_i,\ \mathbf{P}_i \sim MVN(\mathbf{0}_{n_i+1}, \sigma_g^2\mathbf{\Psi}_i),\ \mathbf{E}_i \sim MVN(\mathbf{0}_{n_i+1}, \sigma_e^2\mathbf{I}_{n_i+1}). \quad (1)$$

Here, $\mathbf{\Psi}_i$ 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 $\pi_{ij}$ and that between two relatives $j$ and $j'$ by $\pi_{ijj'}^r$. We let $d_i$ and $d_{ij}^r$ be the inbreeding coefficient for subject $i$ and his/her relative $j$, respectively; the inbreeding coefficient is a parameter quantifying the departure from HWE and ranges from 0 to 1. Then, $\mathbf{\Psi}_i^r$ and $\mathbf{\Psi}_i$ are defined by

$$\mathbf{\Psi}_i^r = \begin{pmatrix} 1+d_{i1}^r & 2\pi_{i12}^r & \cdots & 2\pi_{i1n_i}^r \\ 2\pi_{i12}^r & 1+d_{i2}^r & \ddots & \vdots \\ \vdots & \ddots & \ddots & 2\pi_{i(n_i-1)n_i}^r \\ 2\pi_{i1n_i}^r & \cdots & 2\pi_{i(n_i-1)n_i}^r & 1+d_{in_i}^r \end{pmatrix} \text{ and } \mathbf{\Psi}_i = \begin{pmatrix} 1+d_{i1}^r & \cdots & 2\pi_{i1n_i}^r & 2\pi_{i1} \\ \vdots & \ddots & 2\pi_{i2n_i}^r & \vdots \\ 2\pi_{i1n_i}^r & \cdots & 1+d_{in_i}^r & 2\pi_{in_i} \\ 2\pi_{i1} & \cdots & 2\pi_{in_i} & 1+d_i \end{pmatrix}.$$

Dichotomous phenotypes for subject $i$ and his/her relative $j$ are denoted by $Y_i$ and $Y_{ij}^r$, respectively, and they are coded as 1 for cases and 0 for controls. $Y_i$ and $Y_{ij}^r$ are determined by $L_i$ and $L_{ij}^r$, respectively; if they are larger than a certain threshold, $c$, they become 1, and otherwise they become 0. The phenotype vector for relatives of subject $i$ is denoted by

$$\mathbf{Y}_i^r = \begin{pmatrix} Y_{i1}^r \\ \vdots \\ Y_{in_i}^r \end{pmatrix},$$

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

$$\mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_i^r \\ Y_i \end{pmatrix}.$$

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 $\beta_0$ to be 0 without the loss of generality, and $c$ can be obtained by the following equation:

$$\Phi\left(\frac{-c}{\sqrt{\sigma_g^2+1}}\right)=1-q.$$

If the environmental effect, $Z$, follows the normal distribution, and we denote its variance by $\sigma_z^2$, $c$ can be obtained by

$$\Phi\left(\frac{-c}{\sqrt{\sigma_z^2+\sigma_g^2+1}}\right)=1-q.$$

## Selection of samples with extreme phenotypes

Subjects with extreme phenotypes lead to improvement of statistical power in genetic association studies [19-24], and association analyses have often been conducted with such subjects. At the sample selection stage, genotypes of subjects are not known, and we assume $\beta$ = 0 in equation (1). We can then define the extreme phenotypes for dichotomous phenotypes by the following conditional expectation (CE) of liability scores:

$$\begin{aligned}\text{CE} &= E\left(L_i - \mathrm{Z}_i\beta_0 \mid Y_{i1}^r,\ Y_{i2}^r,\ \ldots,Y_{in_i}^r, Y_i, \mathbf{Z}_i\right) \\ &= E\left(L_i - \mathrm{Z}_i\beta_0 \mid \mathbf{Y}_i, \mathbf{Z}_i\right)\end{aligned}.$$

CEs were calculated with a moment-based method [17], and the detailed algorithm is provided in the Appendix. Once we calculated these for all subjects, $n_a$ affected subjects with the largest CEs and $n_u$ unaffected subjects with the smallest CEs were selected for genetic association studies.

Computation of CEs assumes that $h^2$ (heritability), $q$ (prevalence), $\mathbf{Z}$, and $\beta_0$ 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 $\beta_0$ in 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 [25]. Therefore, if we let

$$\sigma_g^2 = 1.6h^2 \quad \text{and} \quad \sigma_e^2 = 1.6(1-h^2),$$

regression coefficients from logistic regressions can be directly used as $\beta_0$.

**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

$$f_1 = \frac{P(Y_i = 1 | G_i = Dd)}{P(Y_i = 1 | G_i = dd)} \quad \text{and} \quad f_2 = \frac{P(Y_i = 1 | G_i = DD)}{P(Y_i = 1 | G_i = dd)},$$

under HWE, penetrances can be parameterized by

$$P(Y_i = 1 | G_i = dd) = \frac{q}{p_D^2 f_2 + 2 p_D p_d f_1 + p_d^2}, \quad P(Y_i = 1 | G_i = Dd) = P(Y_i = 1 | G_i = dd)\, f_1,$$

$$\text{and} \quad P(Y_i = 1 | G_i = DD) = P(Y_i = 1 | G_i = dd)\, f_2 .$$

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

$$P(G_i | Y_i = 1, \mathbf{Y}_i^r) = \sum_{\mathbf{G}_i^r} P(G_i, \mathbf{G}_i^r | Y_i = 1, \mathbf{Y}_i^r) = \sum_{\mathbf{G}_i^r} \frac{P(Y_i = 1, \mathbf{Y}_i^r | G_i, \mathbf{G}_i^r) P(G_i, \mathbf{G}_i^r)}{P(Y_i = 1, \mathbf{Y}_i^r)} \text{ and}$$

$$P(G_{i'} | Y_{i'} = 0, \mathbf{Y}_{i'}^r) = \sum_{\mathbf{G}_{i'}^r} P(G_{i'}, \mathbf{G}_{i'}^r | Y_{i'} = 0, \mathbf{Y}_{i'}^r) = \sum_{\mathbf{G}_{i'}^r} \frac{P(Y_{i'} = 0, \mathbf{Y}_{i'}^r | G_{i'}, \mathbf{G}_{i'}^r) P(G_{i'}, \mathbf{G}_{i'}^r)}{P(Y_{i'} = 0, \mathbf{Y}_{i'}^r)}.$$

If $\sigma_g^2 = 0$, both conditional probabilities can be simplified to

$$P(G_i \mid Y_i = 1, \mathbf{Y}_i^r) = \frac{P(G_i)P(Y_i = 1 \mid G_i)}{P(Y_i = 1, \mathbf{Y}_i^r)} \sum_{\mathbf{G}_i^r} \left\{ \left( \prod_{j=1}^{n_i} P(Y_{ij}^r \mid G_{ij}^r) \right) P(\mathbf{G}_i^r \mid G_i) \right\},$$

and otherwise, $P(G_i \mid Y_i = 1, \mathbf{Y}_i^r)$ can be numerically calculated. DAFs for case $i$ and control $i'$ can be obtained by

$$P(G_i = DD \mid Y_i = 1, \mathbf{Y}_i^r) + 0.5P(G_i = Dd \mid Y_i = 1, \mathbf{Y}_i^r) \text{ and}$$

$$P(G_{i'} = DD \mid Y_{i'} = 0, \mathbf{Y}_{i'}^r) + 0.5P(G_{i'} = Dd \mid Y_{i'} = 0, \mathbf{Y}_{i'}^r).$$

Therefore, if we assume that there are $n_a$ cases and $n_u$ controls and let

$$p_D^a = \frac{1}{n_a} \sum_{i=1}^{n_a} \left\{ P(G_i = DD \mid Y_i = 1, \mathbf{Y}_i^r) + 0.5P(G_i = Dd \mid Y_i = 1, \mathbf{Y}_i^r) \right\} \text{ and}$$

$$p_D^u = \frac{1}{n_u} \sum_{i'=1}^{n_u} \left\{ P(G_{i'} = DD \mid Y_{i'} = 0, \mathbf{Y}_{i'}^r) + 0.5P(G_{i'} = Dd \mid Y_{i'} = 0, \mathbf{Y}_{i'}^r) \right\},$$

the statistical power for a Cochran Armitage test [26, 27] under the alternative hypothesis can be obtained from

$$\chi^2 \left( \text{df} = 1, \text{NCP} = \frac{(p_D^a - p_D^u)^2}{p_D^a(1 - p_D^a)/n_a + p_D^u(1 - p_D^u)/n_u} \right).$$

If we denote the $\alpha$ quantile of the central chi-square distribution with a single degree of freedom by $\chi_\alpha^2(\text{df} = 1)$, the statistical power at significance level $\alpha$ becomes

$$\text{P}\left\{ \chi^2 \left( \text{df} = 1, \text{NCP} = \frac{(p_D^a - p_D^u)^2}{p_D^a(1 - p_D^a)/n_a + p_D^u(1 - p_D^u)/n_u} \right) > \chi_\alpha^2(\text{df} = 1) \right\}.$$

**Simulation studies**

We assume that there are $n$ subjects with known phenotypes and that $n_a$ cases and $n_u$ controls are selected among these for genotyping ($n \geq n_a + n_u$). 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 1. The $p_D$ was assumed to be 0.2, and genotype frequencies were obtained under HWE. Founders' genotypes in each family were generated from B(2, $p_D$), 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 $\beta_0$ 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 $\sigma_g^2$ and $\sigma_e^2$, respectively. The main genetic effect was obtained by the product of $\beta$ and the number of disease alleles. If we let

$$h^2 = \frac{2\beta^2 p_D p_d + \sigma_g^2}{2\beta^2 p_D p_d + \sigma_g^2 + \sigma_e^2} \text{ and } h_a^2 = \frac{2\beta^2 p_D p_d}{2\beta^2 p_D p_d + \sigma_g^2 + \sigma_e^2},$$

$\sigma_g^2$ and $\beta$ are obtained by the assumed $h^2$ and $h_a^2$. Here, $h^2$ and $h_a^2$ 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 $h_a^2$ to be 0 and 0.005, respectively, and $h^2$ was assumed to be 0.2 and 0.4, respectively. If $h_a^2$ was set to 0, $\beta$ 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 $n_a$ and $n_u$ were assumed to be 500. Genetic association analyses were conducted under the assumption that genotypes were available only for $n_a$ cases and $n_u$ controls.

## 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.

### 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 status 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.

## RESULTS

## Relationship between CEs and disease allele frequencies

Statistical power is positively associated with the difference in DAFs between cases and controls; to investigate any effect of the proposed method on DAF, we assessed the relationship between DAFs and CEs with simulated data. We assumed that $h_a^2 = 0.005$, $h^2 = 0.2$ or 0.4, and $q = 0.1$ or 0.2, and generated 10,000 subjects based on equation (1) under the assumption that there was no environmental effect on phenotype. We then sorted the 10,000 subjects in ascending order of CEs, and subjects were categorized into five equal groups by CE. Figure 2 shows the DAFs according to CE group for cases and controls and indicates that DAFs are proportionally related to CEs. Therefore, we concluded that maximal differences in DAFs between cases and controls could be obtained if affected subjects with the largest DAFs and unaffected subjects with the smallest DAFs were ascertained.

## Evaluation of selection strategy with simulated data

We investigated the effect of the selection strategy with simulated data. 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. Genetic association analyses were conducted with the logistic regression. Empirical type-I errors and power were evaluated for each scenario with 2,000 replicates. Quantile-quantile (QQ) plots (Figure 3) 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 1). Figures 4–5 and Tables 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 $h_a^2 = 0.005$, $h^2 = 0.2$ or 0.4, and $q = 0.1$ or 0.2. Table 4 (scenario 1) shows that S3 was always the most efficient strategy, 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 5 (scenario 2) and Table 6 (scenario 3) show very similar patterns to scenario 1. Therefore, we concluded that cases and controls ascertained with S3 leads to substantial improvement in power.

## 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 $h_a^2 = 0.005$, $h^2= 0.3$, 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 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 6 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.

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

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 [28], and the heritability of T2D has been reported to be approximately 26% [29]. 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 [30]. 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.

The QQ-plots in Figure 7 show that GWAS using all subjects and using only the cases and controls ascertained with S1 and S3 preserve the nominal significance levels, and we concluded therefore that our analyses were statistically valid. Figure 8 shows Manhattan plots for the analyses, with the genome-wide significance level adjusted by Bonferroni correction ($1.872\times10^{-7}$) 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 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.

## DISCUSSION

It has been repeatedly discussed that family history of disease is related to statistical power [8-11]. However, the effect of family history of disease on genetic association analyses has not been carefully investigated, and its use for genetic association analyses has been limited. For instance, affected subjects may be selected for genetic association analyses only if they have at least a certain number of affected relatives [31]. The effect of family history of disease on genetic association analyses is related to both 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 family history of disease. The proposed measures simultaneously take into account both familial distance and number of relatives, and we show that cases and controls ascertained with the proposed method leads to a substantial improvement in power. Our simulation results show that this increase in power should be much larger for common and less heritable diseases. The proposed method was implemented with R code, and it can accept various input file formats such as vcf, PLINK, and gen files. It can be freely downloaded from http://healthstat.snu.ac.kr/software/selSAMPLE.

Furthermore, we showed that DAFs are dependent on the family history of disease, which indicates that ascertainment bias for genetic association analysis can be serious if the family structures are heterogeneous among subjects since it makes DAFs for each family different. It has been also shown that adjustment of heterogeneous ascertainment bias can lead to substantial power improvement for family-based association studies [32, 33], and this effect on statistical power tends to be substantial for highly heritable diseases. The proposed method can be used with minor modifications to adjust the heterogeneous ascertainment bias. For instance, we can calculate CEs for cases and controls with the proposed moment-based methods, and dichotomous phenotypes can be modified with their CEs. The most efficient approach for modifying CEs still requires further investigation.

However, despite the flexibility of the proposed method, there are some limitations. First, our method assumes that the liability scores follow the multivariate normal distribution, and the estimated CEs may be biased if multivariate normality is violated [34]. The generalized linear model can be understood as a latent variable model if its link function is an inverse function of some cumulative distribution [16]. For instance, link functions for logistic and probit regression are inverse functions of the cumulative logistic and standard normal distribution functions, respectively. Therefore, our liability threshold model can be considered an extended probit model [16], and the distribution of unknown liability scores can be chosen by comparing several candidate link functions with Akaike information criteria [35]. Second, there may be recall bias for the family history of disease, and such an effect can be substantial if accuracy is heterogeneous between cases and controls. Third, the proposed method requires that heritability and prevalence are known *a priori*. However, with misspecification of these values, cases and controls ascertained with the proposed method remain the same as long as the order of CEs among affected or unaffected subjects is preserved. In this context, our simulation studies showed that the proposed method is robust against misspecified heritabilities and prevalences for at least the considered scenarios. However, this robustness may be limited to the tested simulation scenarios, and extensive simulation studies are necessary to fully evaluate the sensitivity of the proposed method.

Since high throughput sequencing technology has been introduced, substantial reductions in cost for large-scale genetic association analyses have occurred, and many genetic association analyses have been launched to identify disease susceptibility loci. However, large-scale genetic analyses suffer from serious multiple-testing problems, and sequencing is still often more expensive than phenotyping. Therefore, various statistical methods have been investigated to improve power. Our results reveal that additional statistical power can be achieved in association analyses with carefully selected 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 little cost, and therefore the proposed method may be a useful strategy for the success of genome-wide association analyses.

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF), as funded by the Korean Government [NRF-2014S1A2A2028559]; the Industrial Core Technology Development Program, as funded by the Ministry of Trade, Industry and Energy, Korea (MOTIE) [10040176]; the Basic Science Research Program through the NRF, as funded by the Ministry of Education [NRF-2013R1A1A2010437]; National Institutes of Health [P01 HL105339, R01 HL113264].

## APPENDIX

### Calculation of the conditional expectation (CE)

Conditional expectation (CE) is derived with the moment-based approach with minor modifications [17]. If we let $I_A(\cdot)$ be an indicator function and define that

$$A_i = \begin{cases} (c,\infty) & \text{if } Y_i = 1 \\ (-\infty,c) & \text{if } Y_i = 0 \end{cases},\quad A_{ij}^r = \begin{cases} (c,\infty) & \text{if } Y_{ij}^r = 1 \\ (-\infty,c) & \text{if } Y_{ij}^r = 0 \end{cases},$$

and $\mathbf{I}_{\mathbf{A}_i}(\mathbf{L}_i) = \left( I_{A_{i1}^r}(L_{in_i}^r) \quad \cdots \quad I_{A_{i1}^r}(L_{in_i}^r)\ , I_{A_i}(L_i) \right)'$, the CE for subject $i$ is defined by

$$\mathrm{E}(L_i \mid \mathbf{I}_{\mathbf{A}_i}(\mathbf{L}_i) = \mathbf{1}_{n_i+1}).$$

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 $\mathbf{L}_i$ by

$$f(\mathbf{L}_i) = \left|2\pi\boldsymbol{\Sigma}_i\right|^{-\frac{1}{2}} \exp\left( -\frac{1}{2}\mathbf{L}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{L}_i \right), \text{ where } \operatorname{cov}(\mathbf{L}_i) = \boldsymbol{\Sigma}_i .$$

The conditional pdf of $\mathbf{L}_i$ given $\mathbf{I}_{\mathbf{A}_i}(\mathbf{L}_i) = \mathbf{1}_{n_i+1}$ becomes

$$\begin{aligned} f_{\alpha_i}(\mathbf{L}_i) &= f(\mathbf{L}_i \mid \mathbf{I}_{\mathbf{A}_i}(\mathbf{L}_i) = \mathbf{1}_{n_i+1}) \\ &= \begin{cases} \dfrac{1}{\alpha_i} f(\mathbf{L}_i) & \text{, for } \mathbf{L}_i \in \mathbf{A}_i \\ 0 & \text{, otherwise} \end{cases} \end{aligned}$$

where $\alpha_i = P\left[\mathbf{I}_{\mathbf{A}_i}(\mathbf{L}_i) = \mathbf{1}_{n_i+1}\right]$. We can then find the mgf by

$$\begin{aligned} m(\mathbf{t}_i) &= E\left(e^{\mathbf{t}_i'\mathbf{L}_i} \mid \mathbf{I}_{\mathbf{A}_i}(\mathbf{L}_i) = \mathbf{1}_{n_i+1}\right) \\ &= \frac{1}{\alpha_i \left(2\pi\right)^{(n_i+1)/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} \int_{\mathbf{A}_i} \exp\left\{ -\frac{1}{2}\left( \mathbf{L}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{L}_i - 2\mathbf{t}_i'\mathbf{L}_i \right) \right\} d\mathbf{L}_i \end{aligned}$$

where $\mathbf{t}_i = (t_{i1}^r \ \cdots \ t_{in_i}^r \ , t_i)'$. We let $\boldsymbol{\xi}_i = \boldsymbol{\Sigma}_i\mathbf{t}_i$, and then the exponential term of mgf can be simplified to

$$\exp\left( \frac{1}{2}\mathbf{t}_i'\boldsymbol{\Sigma}_i\mathbf{t}_i \right) \exp\left\{ -\frac{1}{2}\left(\mathbf{L}_i - \boldsymbol{\xi}_i\right)' \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{L}_i - \boldsymbol{\xi}_i\right) \right\},$$

and mgf becomes

$$m\left(\mathbf{t}_i\right)=\frac{\exp\left(\mathbf{t}_i'\boldsymbol{\Sigma}_i\mathbf{t}_i/2\right)}{\alpha_i\left(2\pi\right)^{(n_i+1)/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}}\int_{\mathbf{A}_i}\exp\left(-\frac{1}{2}\mathbf{L}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{L}_i\right)d\mathbf{L}_i.$$

We let $\sigma_{ijk}$ indicate the ($j$,$k$)th element of $\boldsymbol{\Sigma}_i$ and $F_{ik}\left(x\right)$ indicate a marginal pdf for the $k$th element of $\mathbf{L}_i$ of the conditional pdf, $f_{\alpha_i}\left(\mathbf{L}_i\right)$, i.e.,

$$F_{ik}\left(x\right)=\int_{(\mathbf{A}_i)_{-k}}\alpha_i^{-1}f\left((\mathbf{L}_i)_{-k},L_k=x\right)d(\mathbf{L}_i)_{-k}\ ,\ k=1\ ,\ \cdots\ ,\ n_i+1,$$

where subscript –$k$ means that the $k$th element is removed from the corresponding vector. $F_{ik}\left(x\right)$ will be derived in the next section. If we further denote

$$F_{ik}^*=\begin{cases}F_{ik}\left(c\right)-F_{ik}\left(\infty\right), & \text{if } y_{ik}^r=1 \text{ for } k=1\ ,\ \cdots\ ,\ n_i \text{ or } y_i=1 \text{ for } k=n_i+1\\ F_{ik}\left(-\infty\right)-F_{ik}\left(c\right), & \text{otherwise}\end{cases}$$

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

$$\mu_i^*=\left.\frac{\partial m\left(\mathbf{t}_i\right)}{\partial t_i}\right|_{\mathbf{t}_i=\mathbf{0}_{n_i+1}}=\sum_{k=1}^{n_i+1}\sigma_{i(n_i+1)k}F_{ik}^*\ .$$

**Derivation of** $F_{ij}\left(x\right)$

The ($n_i$+1)-dimensional liability vector, $\mathbf{L}_i$, can be partitioned into $(\mathbf{L}_i)_{-j}$ and $L_{ij}^{\ r}$ for $j$ = 1,…,$n_i$ or $\mathbf{L}_i^{\ r}$ and $L_i$ for $j = n_i$+1. For notational convenience, we only considered $j = n_i$+1, which can be readily extended to the other subjects. The partitioned liability vector has the following distribution:

$$\mathbf{L}_i = \begin{pmatrix} \mathbf{L}_i^r \\ L_i \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mathbf{0}_{n_i} \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_i^r & \boldsymbol{\Sigma}_i^{rI} \\ \left(\boldsymbol{\Sigma}_i^{rI}\right)' & 1 \end{pmatrix}\right).$$

If we denote the lower and upper truncated points of $\mathbf{L}_i$ as $\mathbf{a}_i$ and $\mathbf{b}_i$ respectively, the truncated points for $\mathbf{L}_i$ are defined as

$$\mathbf{a}_i = \begin{pmatrix} \mathbf{a}_i^r \\ a_i \end{pmatrix} \text{ and } \mathbf{b}_i = \begin{pmatrix} \mathbf{b}_i^r \\ b_i \end{pmatrix}.$$

When $\mathbf{a}_i < \mathbf{L}_i < \mathbf{b}_i$, the truncated normal distribution function is

$$\begin{aligned} f_\alpha\left(\mathbf{L}_i^r, L_i = x\right) &= \alpha^{-1} f\left(\mathbf{L}_i^r, L_i = x\right) I\left(\mathbf{a}_i < \mathbf{L}_i < \mathbf{b}_i\right) \\ &= \alpha^{-1} f\left(L_i = x\right) f\left(\mathbf{L}_i^r \mid L_i = x\right) I\left(\mathbf{a}_i < \mathbf{L}_i < \mathbf{b}_i\right). \end{aligned}$$

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

$$f\left(L_i = x\right) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Because a conditional distribution of a normal distribution is also normally distributed, we know that $\mathbf{L}_i^{\ r} \mid L_i = x$ is normally distributed with

$$E\left(\mathbf{L}_i^r \mid L_i = x\right) = \boldsymbol{\Sigma}_i^{rI} x \text{ and } \text{var}\left(\mathbf{L}_i^r \mid L_i = x\right) = \boldsymbol{\Sigma}_i^r - \boldsymbol{\Sigma}_i^{rI}\left(\boldsymbol{\Sigma}_i^{rI}\right)'.$$

Therefore, the multivariate marginal pdf of $L_i$ becomes

$$\begin{aligned} F_{i(n_i+1)}\left(x\right) &= \int_{\mathbf{a}_i^r}^{\mathbf{b}_i^r} \alpha^{-1} f\left(L_i = x\right) f\left(\mathbf{L}_i^r \mid L_i = x\right) d\mathbf{L}_i^r \\ &= \alpha^{-1} f\left(L_i = x\right) \int_{\mathbf{a}_i^r}^{\mathbf{b}_i^r} f\left(\mathbf{L}_i^r \mid L_i = x\right) d\mathbf{L}_i^r \end{aligned}$$

Here, $\int_{\mathbf{a}_i^r}^{\mathbf{b}_i^r} f\left(\mathbf{L}_i^r \mid L_i = x\right) d\mathbf{L}_i^r$ can be computed using statistical software, such as the function pmvnorm() in the R package mvtnorm.

## References

1. Mardis ER: **The impact of next-generation sequencing technology on genetics**. *Trends in genetics : TIG* 2008, **24**(3):133-141.
2. Metzker ML: **Sequencing technologies - the next generation**. *Nature reviews Genetics* 2010, **11**(1):31-46.
3. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB: **The real cost of sequencing: higher than you think!** *Genome Biol* 2011, **12**(8).
4. Moore GE: **Cramming more components onto integrated circuits (Reprinted from Electronics, pg 114-117, April 19, 1965)**. *P Ieee* 1998, **86**(1):82-85.
5. Maher B: **Personal genomes: The case of the missing heritability**. *Nature* 2008, **456**(7218):18-21.
6. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A *et al*: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**(7265):747-753.
7. Pritchard JK: **Are rare variants responsible for susceptibility to complex diseases?** *The American Journal of Human Genetics* 2001, **69**(1):124-137.
8. Antoniou AC, Easton DF: **Polygenic inheritance of breast cancer: Implications for design of association studies**. *Genetic epidemiology* 2003, **25**(3):190-202.
9. Howson JM, Barratt BJ, Todd JA, Cordell HJ: **Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci**. *Genetic epidemiology* 2005, **29**(1):51-67.
10. Li M, Boehnke M, Abecasis GR: **Efficient study designs for test of genetic association using sibship data and unrelated cases and controls**. *American journal of human genetics* 2006, **78**(5):778-792.
11. Risch N: **Implications of multilocus inheritance for gene-disease association studies**. *Theoretical population biology* 2001, **60**(3):215-220.
12. Edwards J: **Familial predisposition in man**. *British Medical Bulletin* 1969, **25**(1):58-64.
13. Falconer DS: **The inheritance of liability to certain diseases, estimated from the incidence among relatives**. *Annals of Human Genetics* 1965, **29**(1):51-76.
14. Stroup WW: **Generalized linear mixed models: modern concepts, methods and applications**: CRC press; 2012.
15. Lee SH, Wray NR, Goddard ME, Visscher PM: **Estimating missing heritability for disease from genome-wide association studies**. *The American Journal of Human Genetics* 2011, **88**(3):294-305.
16. Bliss C: **The method of probits**. *Science* 1934, **79**(2037):38-39.
17. Wilhelm S: **Moments Calculation For the Doubly Truncated Multivariate Normal Density**. *arXiv preprint arXiv:12065387* 2012.
18. Falconer DS: **The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus**. *Annals of human genetics* 1967, **31**(1):1-20.
19. Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, Lyssenko V, Jonsson A, Lindholm E, Tuomi T, Isomaa B: **Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants**. *Genetic epidemiology* 2011, **35**(4):236-246.
20. Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D: **Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies**. *Genetic epidemiology* 2011, **35**(8):790-799.
21. Barnett IJ, Lee S, Lin X: **Detecting rare variant effects using extreme phenotype**

**sampling in sequencing association studies**. *Genetic epidemiology* 2013, **37**(2):142-151.

22. Nebert DW: **Extreme discordant phenotype methodology: an intuitive approach to clinical pharmacogenetics**. *European journal of pharmacology* 2000, **410**(2):107-120.
23. Perez-Gracia JL, Ruiz-Ilundain MG, Garcia-Ribas I, Carrasco EM: **The role of extreme phenotype selection studies in the identification of clinically relevant genotypes in cancer research**. *CANCER-PHILADELPHIA-* 2002, **95**(7):1605-1610.
24. Risch NJ, Zhang H: **Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations**. *American journal of human genetics* 1996, **58**(4):836.
25. Gelman A, Hill J: **Data analysis using regression and multilevel/hierarchical models**: Cambridge University Press; 2006.
26. Cochran WG: **Some Methods for Strengthening the Common &#x3c7;<sup>2</sup> Tests**. *Biometrics* 1954, **10**(4):417-451.
27. Armitage P: **Tests for linear trends in proportions and frequencies**. *Biometrics* 1955, **11**(3):375-386.
28. Kim DJ: **The epidemiology of diabetes in Korea**. *Diabetes & metabolism journal* 2011, **35**(4):303-308.
29. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H: **Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study**. *Diabetologia* 1999, **42**(2):139-145.
30. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nature genetics* 2006, **38**(8):904-909.
31. Risch N, Teng J: **The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling**. *Genome Research* 1998, **8**(12):1273-1288.
32. Zaitlen N, Loh P-R, Vilhjalmsson B, Pollack S, Gusev A, Yang J, Chen G-B, Goddard ME, Visscher PM, Patterson N: **Mixed Model with Correction for Case-Control Ascertainment Increases Association Power**. *bioRxiv* 2014:008755.
33. Park S, Lee S, Lee Y, Herold C, Hooli B, Mullin K, Park T, Park C, Bertram L, Lange C: **Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families**. *BMC medical genetics* 2015, **16**(1):62.
34. Benchek PH, Morris NJ: **How meaningful are heritability estimates of liability?** *Human genetics* 2013, **132**(12):1351-1360.
35. Bozdogan H: **Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions**. *Psychometrika* 1987, **52**(3):345-370.

**Table 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 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.1 | 0.005 | 0.0055 | 0.0065 | 0.0040 | 0.0070 | 0.0050 |
| | | 0.01 | 0.0070 | 0.0135 | 0.0090 | 0.0100 | 0.0105 |
| | | 0.05 | 0.0515 | 0.0605 | 0.0510 | 0.0525 | 0.0555 |
| | 0.2 | 0.005 | 0.0020 | 0.0050 | 0.0040 | 0.0070 | 0.0070 |
| | | 0.01 | 0.0050 | 0.0090 | 0.0100 | 0.0110 | 0.0115 |
| | | 0.05 | 0.0395 | 0.0430 | 0.0550 | 0.0540 | 0.0520 |
| 0.4 | 0.1 | 0.005 | 0.0045 | 0.0045 | 0.0050 | 0.0040 | 0.0060 |
| | | 0.01 | 0.0090 | 0.0120 | 0.0115 | 0.0085 | 0.0145 |
| | | 0.05 | 0.0440 | 0.0475 | 0.0450 | 0.0445 | 0.0495 |
| | 0.2 | 0.005 | 0.0050 | 0.0050 | 0.0045 | 0.0035 | 0.0070 |
| | | 0.01 | 0.0110 | 0.0095 | 0.0085 | 0.0085 | 0.0105 |
| | | 0.05 | 0.0555 | 0.0490 | 0.0460 | 0.0470 | 0.0510 |

**Table 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 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.1 | 0.005 | 0.0035 | 0.0035 | 0.0040 | 0.0040 | 0.0040 |
| | | 0.01 | 0.0075 | 0.0095 | 0.0090 | 0.0095 | 0.0105 |
| | | 0.05 | 0.0500 | 0.0560 | 0.0500 | 0.0500 | 0.0500 |
| | 0.2 | 0.005 | 0.0070 | 0.0030 | 0.0050 | 0.0065 | 0.0065 |
| | | 0.01 | 0.0145 | 0.0095 | 0.0080 | 0.0095 | 0.0090 |
| | | 0.05 | 0.0545 | 0.0415 | 0.0455 | 0.0460 | 0.0535 |
| 0.4 | 0.1 | 0.005 | 0.0055 | 0.0090 | 0.0075 | 0.0045 | 0.0035 |
| | | 0.01 | 0.0100 | 0.0155 | 0.0120 | 0.0090 | 0.0095 |
| | | 0.05 | 0.0455 | 0.0555 | 0.0520 | 0.0420 | 0.0440 |
| | 0.2 | 0.005 | 0.0070 | 0.0050 | 0.0030 | 0.0035 | 0.0055 |
| | | 0.01 | 0.0130 | 0.0100 | 0.0075 | 0.0065 | 0.0110 |
| | | 0.05 | 0.0530 | 0.0570 | 0.0535 | 0.0500 | 0.0475 |

**Table 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 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.1 | 0.005 | 0.0050 | 0.0045 | 0.0030 | 0.0025 | 0.0035 |
| | | 0.01 | 0.0070 | 0.0090 | 0.0080 | 0.0085 | 0.0085 |
| | | 0.05 | 0.0470 | 0.0450 | 0.0580 | 0.0525 | 0.0515 |
| | 0.2 | 0.005 | 0.0040 | 0.0055 | 0.0060 | 0.0070 | 0.0065 |
| | | 0.01 | 0.0075 | 0.0090 | 0.0105 | 0.0120 | 0.0135 |
| | | 0.05 | 0.0420 | 0.0440 | 0.0570 | 0.0570 | 0.0495 |
| 0.4 | 0.1 | 0.005 | 0.0060 | 0.0075 | 0.0055 | 0.0025 | 0.0050 |
| | | 0.01 | 0.0095 | 0.0135 | 0.0105 | 0.0095 | 0.0115 |
| | | 0.05 | 0.0450 | 0.0560 | 0.0480 | 0.0500 | 0.0515 |
| | 0.2 | 0.005 | 0.0055 | 0.0040 | 0.0060 | 0.0040 | 0.0045 |
| | | 0.01 | 0.0085 | 0.0075 | 0.0120 | 0.0080 | 0.0085 |
| | | 0.05 | 0.0475 | 0.0450 | 0.0460 | 0.0480 | 0.0455 |

**Table 4. Empirical power estimates for scenario 1.** The empirical power levels were estimated with 2,000 replicates at several significance levels. We assume that $h_a^2$=0.005, $h^2$ = 0.2 and 0.4, and $q$ = 0.1 and 0.2.

| $h^2$ | $q$ | Significance levels | S1 | S2 | **S3** | S4 | S5 |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.1 | 0.005 | 0.2675 | 0.4820 | **0.6635** | 0.4255 | 0.0030 |
| | | 0.01 | 0.3505 | 0.5795 | **0.7450** | 0.5245 | 0.0085 |
| | | 0.05 | 0.5880 | 0.8070 | **0.8980** | 0.7545 | 0.0520 |
| | 0.2 | 0.005 | 0.2210 | 0.5520 | **0.8220** | 0.4825 | 0.0095 |
| | | 0.01 | 0.2840 | 0.6515 | **0.8815** | 0.5745 | 0.0195 |
| | | 0.05 | 0.5260 | 0.8480 | **0.9645** | 0.7790 | 0.0930 |
| 0.4 | 0.1 | 0.005 | 0.2700 | 0.4445 | **0.6090** | 0.4325 | 0.0085 |
| | | 0.01 | 0.3525 | 0.5285 | **0.6925** | 0.5130 | 0.0155 |
| | | 0.05 | 0.5950 | 0.7640 | **0.8670** | 0.7530 | 0.0675 |
| | 0.2 | 0.005 | 0.1825 | 0.4730 | **0.7010** | 0.4210 | 0.0055 |
| | | 0.01 | 0.2425 | 0.5625 | **0.7825** | 0.5005 | 0.0135 |
| | | 0.05 | 0.4725 | 0.7855 | **0.9215** | 0.7210 | 0.0530 |

**Table 5. Empirical power estimates for scenario 2.** The empirical power levels were estimated with 2,000 replicates at several significance levels. We assume that $h_a^2$=0.005, $h^2$ = 0.2 and 0.4, and $q$ = 0.1 and 0.2.

| $h^2$ | $q$ | Significance levels | S1 | S2 | **S3** | S4 | S5 |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.1 | 0.005 | 0.2715 | 0.4960 | **0.7275** | 0.5165 | 0.0070 |
| | | 0.01 | 0.3555 | 0.5855 | **0.7970** | 0.6160 | 0.0110 |
| | | 0.05 | 0.6115 | 0.8010 | **0.9320** | 0.8240 | 0.0415 |
| | 0.2 | 0.005 | 0.1930 | 0.5940 | **0.9000** | 0.5485 | 0.0165 |
| | | 0.01 | 0.2750 | 0.6840 | **0.9310** | 0.6530 | 0.0270 |
| | | 0.05 | 0.5030 | 0.8595 | **0.9775** | 0.8415 | 0.0960 |
| 0.4 | 0.1 | 0.005 | 0.2630 | 0.4355 | **0.6425** | 0.4625 | 0.0060 |
| | | 0.01 | 0.3540 | 0.5285 | **0.7320** | 0.5585 | 0.0120 |
| | | 0.05 | 0.5955 | 0.7495 | **0.8930** | 0.7875 | 0.0555 |
| | 0.2 | 0.005 | 0.1910 | 0.5080 | **0.7940** | 0.4870 | 0.0050 |
| | | 0.01 | 0.2695 | 0.5975 | **0.8520** | 0.5800 | 0.0080 |
| | | 0.05 | 0.4985 | 0.8030 | **0.9525** | 0.7885 | 0.0480 |

**Table 6. Empirical power estimates for scenario 3.** The empirical power levels were estimated with 2,000 replicates at several significance levels. We assume that $h_a^2$=0.005, $h^2$ = 0.2 and 0.4, and $q$ = 0.1 and 0.2.

| $h^2$ | $q$ | Significance levels | S1 | S2 | **S3** | S4 | S5 |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.1 | 0.005 | 0.2700 | 0.4970 | **0.7475** | 0.5180 | 0.0045 |
| | | 0.01 | 0.3490 | 0.5825 | **0.8065** | 0.6075 | 0.0095 |
| | | 0.05 | 0.5980 | 0.7950 | **0.9245** | 0.8120 | 0.0405 |
| | 0.2 | 0.005 | 0.2135 | 0.5635 | **0.8860** | 0.5770 | 0.0185 |
| | | 0.01 | 0.2850 | 0.6505 | **0.9215** | 0.6595 | 0.0340 |
| | | 0.05 | 0.5380 | 0.8385 | **0.9825** | 0.8565 | 0.1130 |
| 0.4 | 0.1 | 0.005 | 0.2615 | 0.4455 | **0.6375** | 0.4470 | 0.0090 |
| | | 0.01 | 0.3485 | 0.5330 | **0.7205** | 0.5390 | 0.0185 |
| | | 0.05 | 0.5855 | 0.7570 | **0.8795** | 0.7710 | 0.0655 |
| | 0.2 | 0.005 | 0.2130 | 0.4695 | **0.7860** | 0.5025 | 0.0090 |
| | | 0.01 | 0.2890 | 0.5775 | **0.8475** | 0.6005 | 0.0175 |
| | | 0.05 | 0.5020 | 0.7890 | **0.9515** | 0.7990 | 0.0570 |

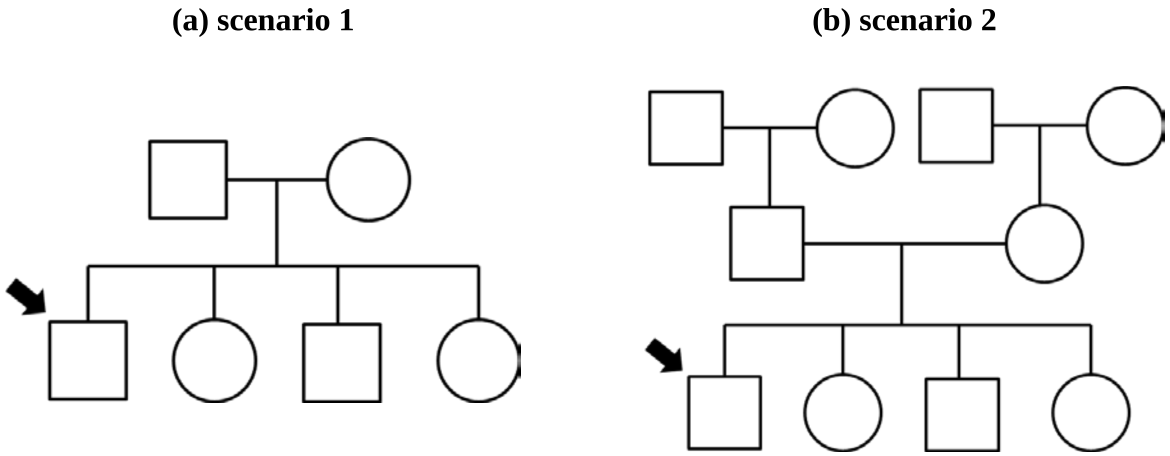
**Table 7. Empirical relative power estimates for misspecified heritabilities and prevalences for scenario 3.** The empirical power levels were estimated with 2,000 replicates at several significance levels and the ratios of the power estimates from misspecified ($h^2$, $q$) to those from the correctly defined ($h^2$, $q$) were calculated in percentage. We assume that $h_a^2$=0.005, and ($h^2$, $q$) = (0.3, 0.3) for generating phenotypes, and four misspecified pairs of ($h^2$, $q$) were considered.

| $h^2$ | $q$ | Significance levels | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|
| 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 | 0.4 | 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 | 0.5 | 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 |

**Table 8. Results from GWAS.** The significance level adjusted by Bonferroni correction is $1.872\times10^{-7}$ 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 | **$8.25\times10^{-19}$** | **$2.03\times10^{-9}$** | **$3.35\times10^{-15}$** |
| rs7754840 | 6 | 20661250 | CDKAL1 | **$7.03\times10^{-17}$** | **$1.82\times10^{-8}$** | **$1.88\times10^{-12}$** |
| rs9460546 | 6 | 20663632 | CDKAL1 | **$5.10\times10^{-16}$** | **$6.53\times10^{-8}$** | **$3.91\times10^{-12}$** |
| rs9465871 | 6 | 20717255 | CDKAL1 | **$8.91\times10^{-16}$** | $2.40\times10^{-7}$ | **$1.61\times10^{-11}$** |
| rs7747752 | 6 | 20725423 | CDKAL1 | **$1.31\times10^{-15}$** | **$1.69\times10^{-7}$** | **$5.39\times10^{-12}$** |
| rs7767391 | 6 | 20725240 | CDKAL1 | **$1.84\times10^{-15}$** | **$1.78\times10^{-7}$** | **$7.21\times10^{-12}$** |
| rs9348440 | 6 | 20641336 | CDKAL1 | **$1.20\times10^{-14}$** | $5.90\times10^{-7}$ | **$3.35\times10^{-11}$** |
| rs2328549 | 6 | 20718240 | CDKAL1 | **$3.53\times10^{-14}$** | $2.48\times10^{-6}$ | **$5.02\times10^{-11}$** |
| rs2328529 | 6 | 20631953 | CDKAL1 | **$5.52\times10^{-10}$** | $3.35\times10^{-6}$ | $4.34\times10^{-7}$ |
| rs10811661 | 9 | 22134094 | CDKN2B-AS1 | **$2.84\times10^{-9}$** | **$1.51\times10^{-8}$** | $1.04\times10^{-6}$ |
| rs7741604 | 6 | 20731524 | CDKAL1 | **$4.74\times10^{-9}$** | $1.16\times10^{-5}$ | $2.23\times10^{-6}$ |
| rs1526959 | 12 | 79753790 | SYT1 | **$1.16\times10^{-8}$** | $3.00\times10^{-3}$ | $2.89\times10^{-6}$ |
| rs4291090 | 6 | 20570039 | CDKAL1 | **$1.81\times10^{-8}$** | $3.20\times10^{-4}$ | $6.40\times10^{-7}$ |
| rs2820001 | 6 | 20758943 | CDKAL1 | **$3.23\times10^{-8}$** | $9.19\times10^{-5}$ | $2.05\times10^{-5}$ |
| rs10946406 | 6 | 20758760 | CDKAL1 | **$4.01\times10^{-8}$** | $1.61\times10^{-2}$ | $5.02\times10^{-7}$ |
| rs2294809 | 6 | 20599888 | CDKAL1 | **$4.52\times10^{-8}$** | $4.90\times10^{-4}$ | $2.41\times10^{-6}$ |
| rs9366357 | 6 | 20599628 | CDKAL1 | **$6.09\times10^{-8}$** | $4.34\times10^{-4}$ | $4.22\times10^{-6}$ |
| rs12679402 | 8 | 41958980 | AP3M2 | $8.45\times10^{-5}$ | $2.53\times10^{-3}$ | **$1.26\times10^{-8}$** |

**Figure 1. Family history of disease.** The person indicated by an arrow is a proband.



**(a) scenario 1**

**(b) scenario 2**

**Figure 2. DAFs according to CEs.** Figure 2A and Figure 2B shows DAFs for cases and controls respectively. All subjects were sorted with CEs and classified to 5 different groups with CEs.

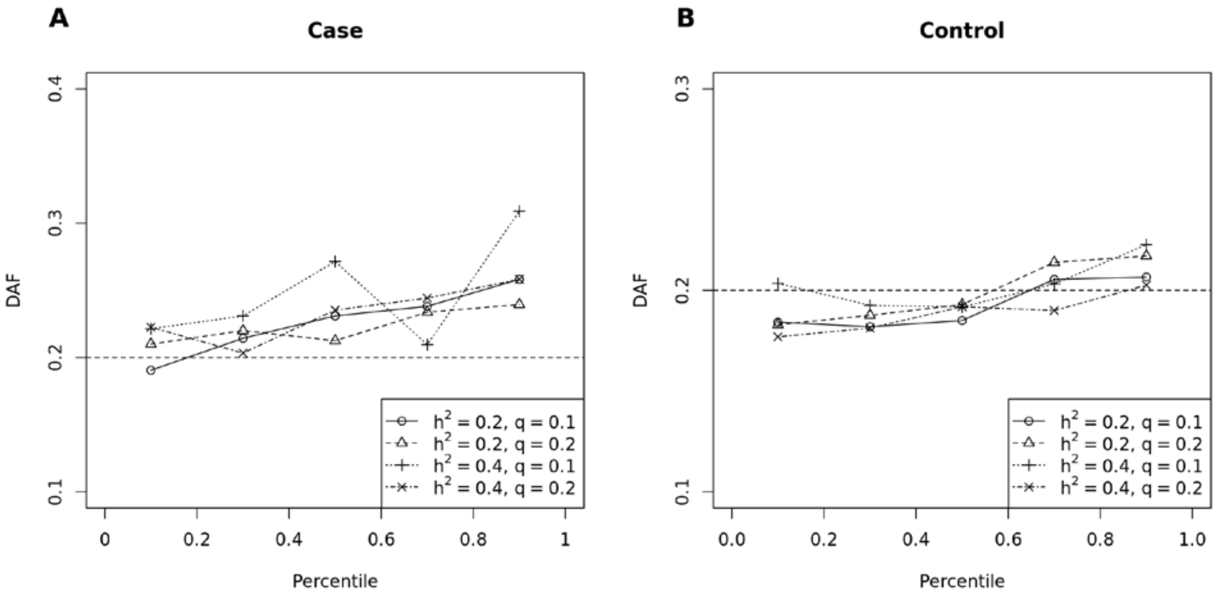


**Figure 3. QQ-plots of simulated data for scenario 1.** We assume that $h^2$=0.2 and $q$ = 0.1 and Scenario 1 was assumed for relatives' family structure. QQ-plots were generated from 2,000 replicates.

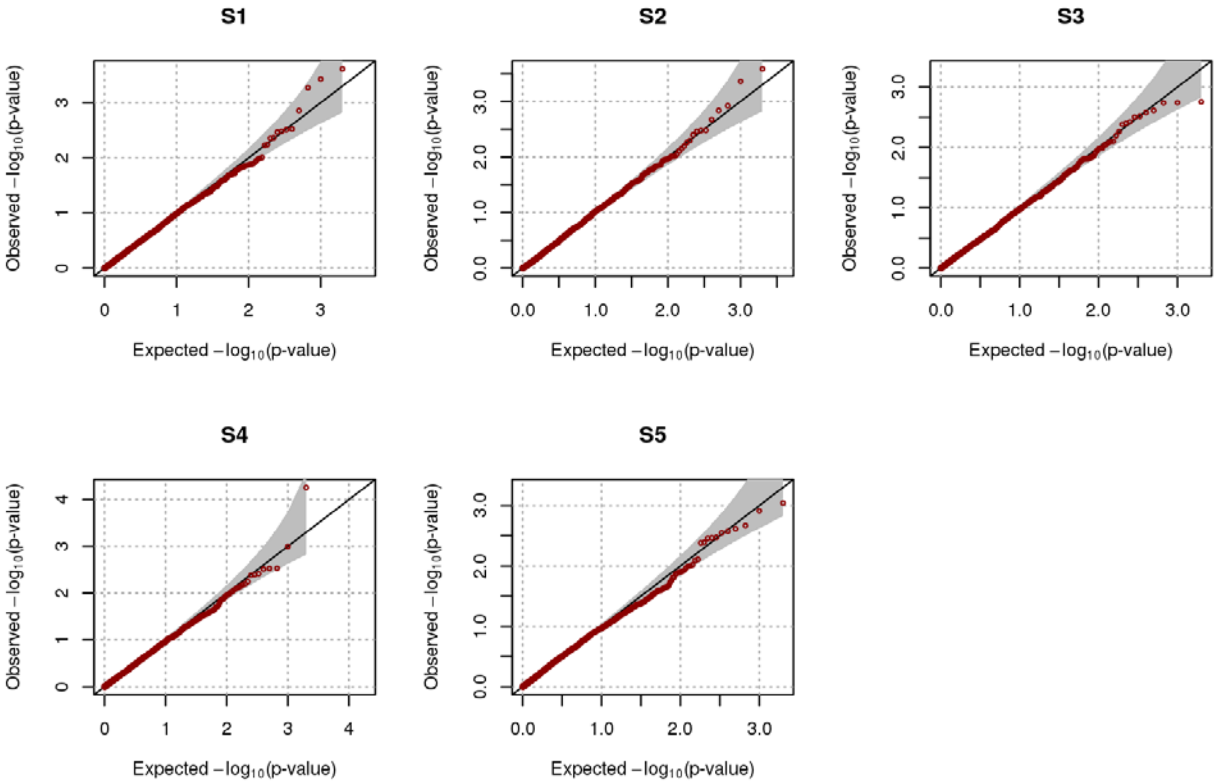


**Figure 4. QQ-plots of simulated data for scenario 2.** We assume that $h^2$=0.2 and $q$ = 0.1 and Scenario 2 was assumed for relatives' family structure. QQ-plots were generated from 2,000 replicates.

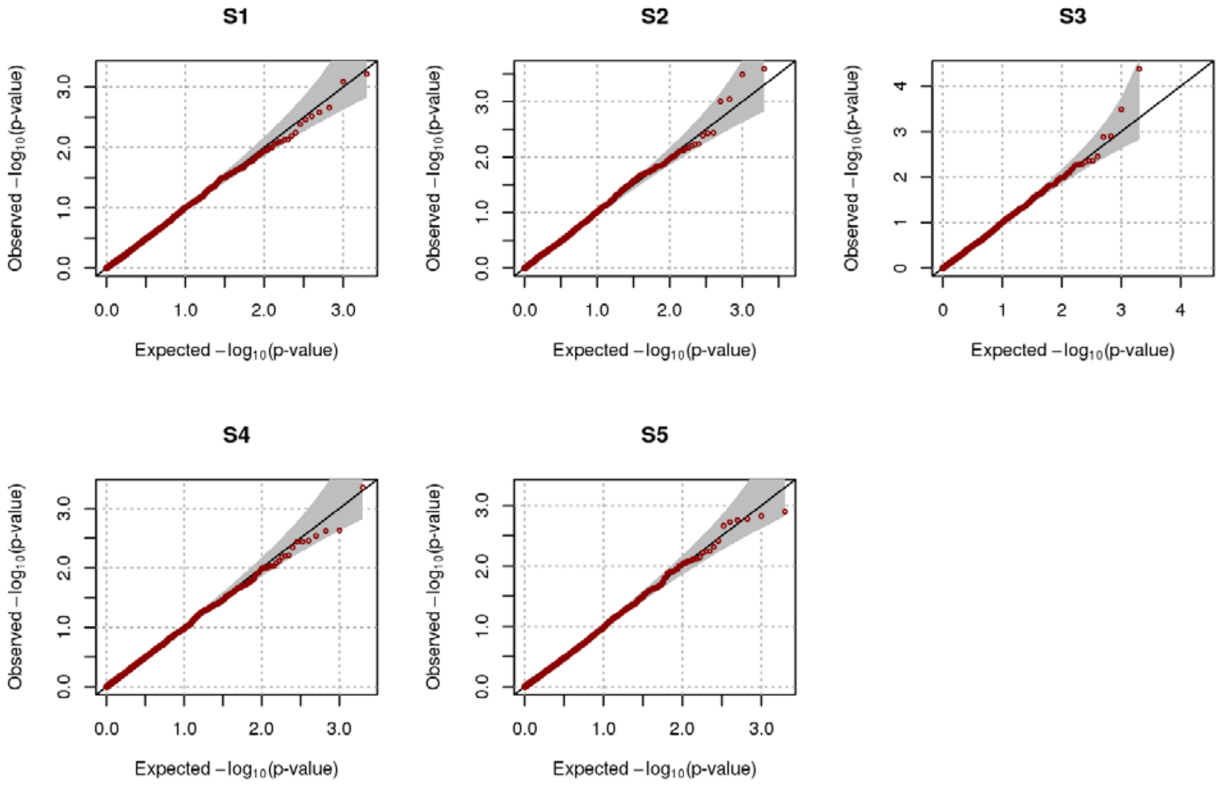


**Figure 5. QQ-plots of simulated data for scenario 3.** We assume that $h^2$=0.2 and $q$ = 0.1 and Scenario 3 was assumed for relatives' family structure. QQ-plots were generated from 2,000 replicates.

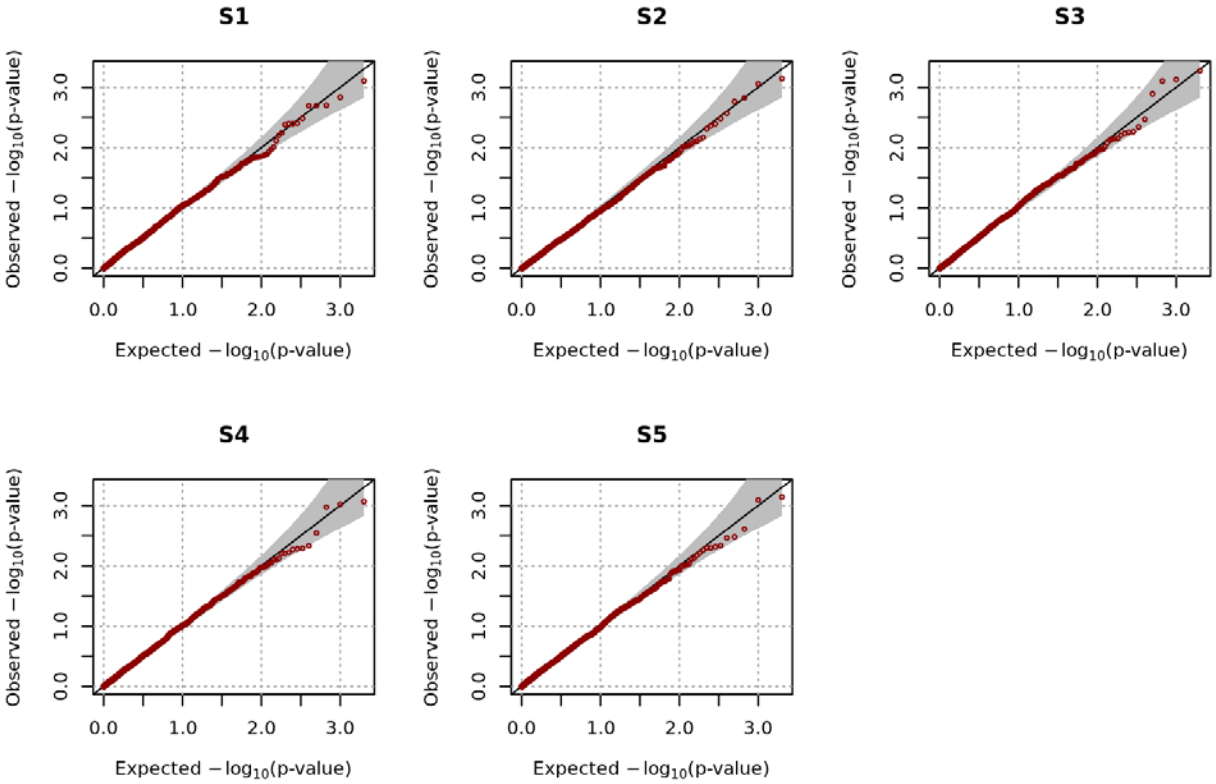


**Figure 6. Contour plot for the correlation between orders of CEs calculated from true and misspecified ($h^2$, $q$).** Orders of CEs were obtained for the various choices of heritability and prevalence and their correlations with true orders were calculated. Data was generated from ($h^2$, $q$) = (0.3, 0.3) and ' **x** ' is a point where correlation is exactly 1.

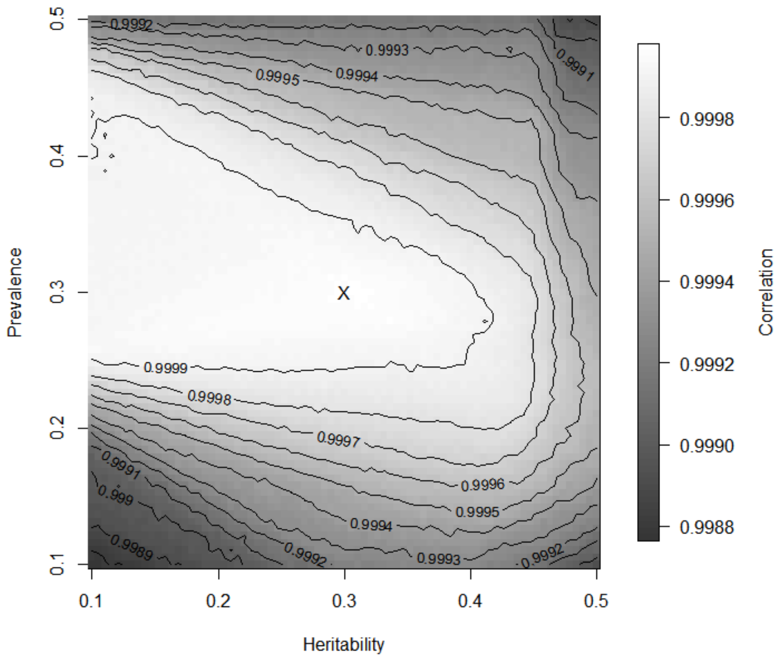


**Figure 7. QQ-plots for the results from GWAS of T2D**.

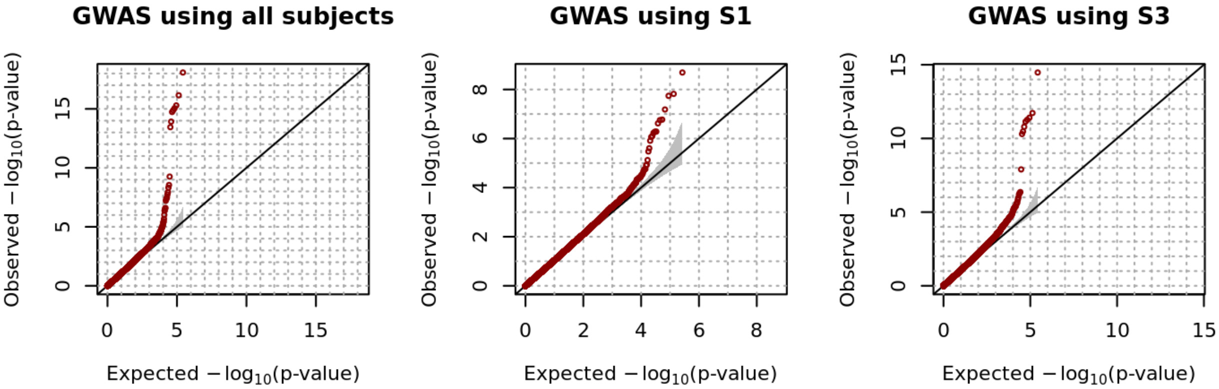
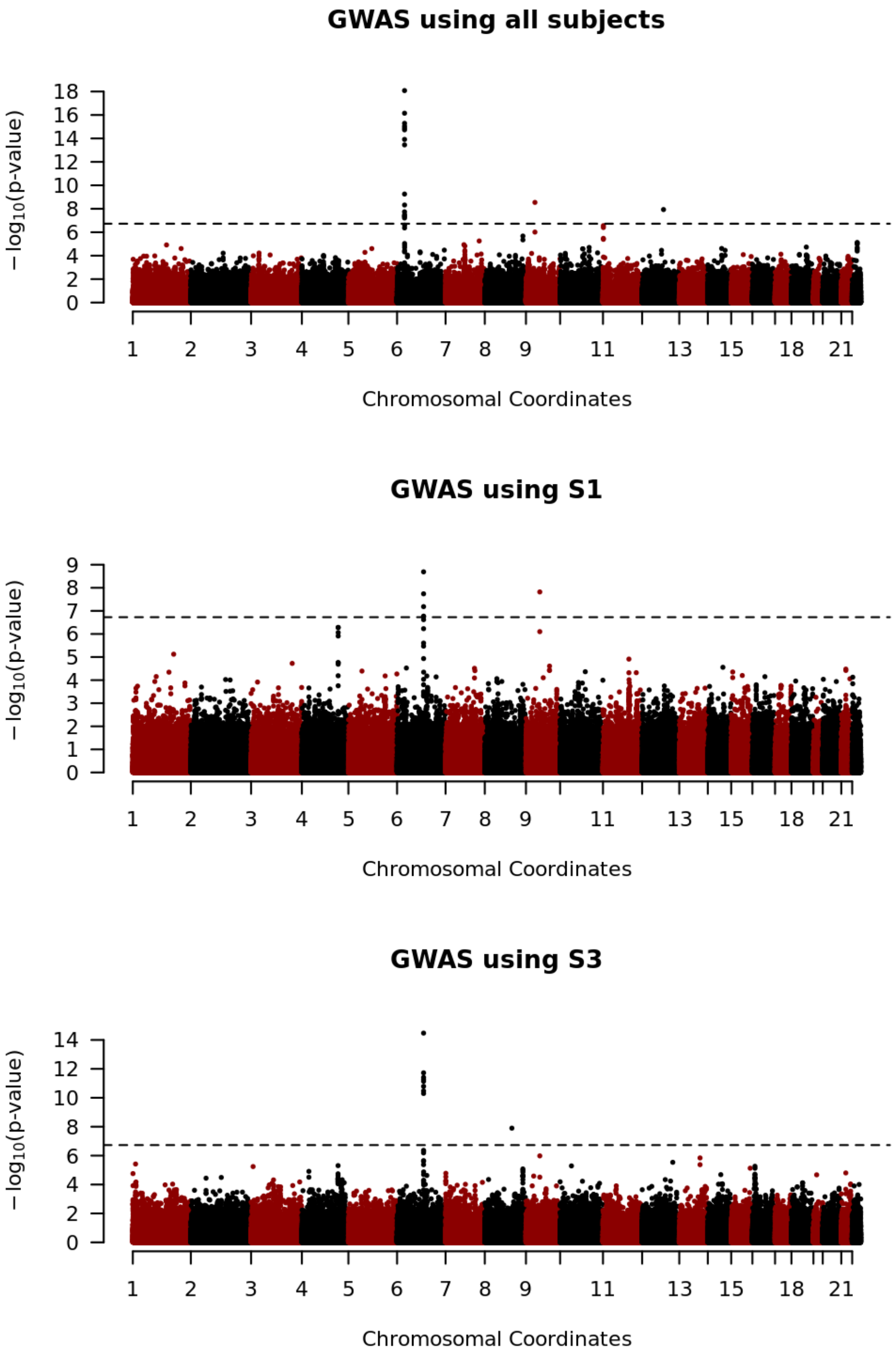


**Figure 8. Manhattan-plots for the results from GWAS of T2D.**



GWAS using all subjects

GWAS using S1

GWAS using S3