Published in final edited form as:

Stat Med. 2017 June 15; 36(13): 2081–2099. doi:10.1002/sim.7248.

Selecting cases and controls for DNA sequencing studies using family histories of disease

Wonji Kim^a, Dandi Qiao^b, Michael H. Cho^{b,c}, Soo Heon Kwak^d, Kyong Soo Park^d, Edwin K. Silverman^{b,c}, Pak Sham^{e,f,g}, and Sungho Won^{a,h,i,j,*†}

^aInterdisciplinary Program of Bioinformatics, Seoul National University, Seoul, Korea

^bChanning Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA

^cDivision of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, 02115, USA

^dDepartment of Internal Medicine, Seoul National University College of Medicine, Seoul, Korea

^eDepartment of Psychiatry, University of Hong Kong, Hong Kong, SAR, China

^fGenome Research Centre, University of Hong Kong, Hong Kong, SAR, China

^gState Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong, SAR, China

^hDepartment of Public Health Science, Seoul National University, Seoul, Korea

ⁱInstitute of Health and Environment, Seoul National University, Seoul, Korea

^jNational Cancer Center, Seoul, Korea

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 family history. We assume that disease status is determined by unobserved liability scores. Our method consists of two steps: first, the conditional means of liability are estimated with the liability threshold model given the individual's disease status and those of their relatives. 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

*Correspondence to: Sungho Won, Department of Public Health Science, Seoul National University, 1 Kwanak-ro Kwanak-gu, Seoul 151-742, Korea.

†won1@snu.ac.kr

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.

applied to genome-wide association studies for type 2 diabetes, and our analysis results reveal the practical value of the proposed methods.

Keywords

family history of disease; risk prediction; liability

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 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 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 has 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 because of 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 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 versus 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 (CE) 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 risks are selected as cases and controls, respectively. The proposed methods were applied to genome-wide association studies (GWASs) for type 2 diabetes (T2D) with data collected from the Korea Association Resource (KARE) project and Seoul National University Hospital (SNUH) in Korea. 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 ($i = 1, \dots, n$). 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 the Hardy–Weinberg equilibrium (HWE) in a population. We denote the genotypes of subject i and his or her relative j by G_i and G_{ij}^r , respectively, and the genotype vectors are defined by

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

We consider the liability threshold model [13], and dichotomous phenotypes are determined by the unobserved continuous liability score. The liability scores of subject i and his or 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 or 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 in which all elements are 0 and 1 by $\mathbf{0}_w$ and $\mathbf{1}_w$, respectively.

Then, if we let σ_g^2 and σ_e^2 be variances of polygenic effects and random residual effects, respectively, and let $\mathbf{Z}_i$ include the intercept, we can assume that

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

Here, Ψ_i indicates the kinship coefficient matrix for both subject i and his or her relatives. We denote the kinship coefficient between subject i and his or her relative j by π_{ij} and that between two relatives j and j' by $\pi_{ijj'}^r$. Similarly, d_i and d_{ij}^r denote the inbreeding coefficients for subject i and his or 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, for example, *pedigreemm* [18,19]. Then, Ψ_i^r and Ψ_i are defined by

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

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 or 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. In a liability threshold model, Y_i and Y_{ij}^r are determined by L_i and L_{ij}^r , respectively; if L_i and L_{ij}^r are above a certain threshold value c , Y_i and Y_{ij}^r 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

$$\mathbf{Y}_i^r = \begin{pmatrix} Y_{i1}^r \\ \vdots \\ Y_{in_i}^r \end{pmatrix} = \begin{pmatrix} I(L_{i1}^r > c) \\ \vdots \\ I(L_{in_i}^r > c) \end{pmatrix},$$

and that for the subject i and his or 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 β_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 σ_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 [20–25], 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). 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 CE of liability scores:

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

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 β_0 are known. While h^2 , q , and $\mathbf{Z}$ 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_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 [26]. Therefore, if we let

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

regression coefficients from logistic regressions can be directly used as β_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

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

under HWE, penetrances can be parameterized by

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

and $P(Y_i=1|G_i=DD) = P(Y_i=1|G_i=dd) f_2$.

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

$$\begin{aligned} P(G_i|Y_i=1, \mathbf{Y}_i^r) &= \sum_{\mathbf{G}_i^r} P(G_i, \mathbf{G}_i^r | Y_i=1, \mathbf{Y}_i^r) \\ &= \sum_{\mathbf{G}_i^r} \frac{P(Y_i=1, \mathbf{Y}_i^r | G_i, \mathbf{G}_i^r) P(G_i, \mathbf{G}_i^r)}{P(Y_i=1, \mathbf{Y}_i^r)} \text{ and } P(G_{i'}|Y_{i'}=0, \mathbf{Y}_{i'}^r) \\ &= \sum_{\mathbf{G}_{i'}^r} P(G_{i'}, \mathbf{G}_{i'}^r | Y_{i'}=0, \mathbf{Y}_{i'}^r) = \sum_{\mathbf{G}_{i'}^r} \frac{P(Y_{i'}=0, \mathbf{Y}_{i'}^r | G_{i'}, \mathbf{G}_{i'}^r) P(G_{i'}, \mathbf{G}_{i'}^r)}{P(Y_{i'}=0, \mathbf{Y}_{i'}^r)}. \end{aligned}$$

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

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

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

$$\begin{aligned} &P(G_i=DD|Y_i=1, \mathbf{Y}_i^r) + 0.5P(G_i=Dd|Y_i=1, \mathbf{Y}_i^r) \text{ and} \\ &P(G_{i'}=DD|Y_{i'}=0, \mathbf{Y}_{i'}^r) + 0.5P(G_{i'}=Dd|Y_{i'}=0, \mathbf{Y}_{i'}^r). \end{aligned}$$

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

$$\begin{aligned} p_D^a &= \frac{1}{n_a} \sum_{i=1}^{n_a} \{P(G_i=DD|Y_i=1, \mathbf{Y}_i^r) + 0.5P(G_i=Dd|Y_i=1, \mathbf{Y}_i^r)\} \text{ and} \\ p_D^u &= \frac{1}{n_u} \sum_{i'=1}^{n_u} \{P(G_{i'}=DD|Y_{i'}=0, \mathbf{Y}_{i'}^r) + 0.5P(G_{i'}=Dd|Y_{i'}=0, \mathbf{Y}_{i'}^r)\}, \end{aligned}$$

the statistical power for a Cochran–Armitage test [27,28] under the alternative hypothesis can be obtained from

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

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

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

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 ($n = n_a + n_u$). 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 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 σ_g^2 and σ_e^2 , respectively. The main genetic effect was obtained by the product of β and the number of disease alleles. If we let

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

σ_g^2 and β 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, β became 0, which indicates the null hypothesis (no association between genetic variants and phenotypes). Empirical size and power estimates were calculated with 2000 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.

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.

The Korea Association Resource 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 8842 participants consisting of 4183 men and 4659 women, 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, 1179 subjects were diagnosed as having T2D by a standard guideline (glucose at baseline ≥ 126 mg/dL, glucose 120 min after the insulin challenge ≥ 200 mg/dL, or HbA1c $\geq 6.5\%$). The disease status of their relatives was collected by a survey from all participants, and 1037 subjects (125 cases and 912 controls) answered that they have affected relatives. In total, there were 1230 affected relatives available.

The 8842 subjects were genotyped for 352,228 single-nucleotide polymorphisms (SNPs) with the Affymetrix Genome-Wide Human SNP Array 6.0. In our GWASs, 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 were less than 0.05. We also eliminated subjects with gender inconsistencies, whose identity by state was more than 0.8 or whose call rates were less than 95%. As a result, 310,515 SNPs for 8842 subjects were utilized for GWAS.

The Seoul National University Hospital data

Type 2 diabetes patients were diagnosed by World Health Organization criteria from 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 7825 relatives were available, among which 2875 subjects had T2D.

Type 2 diabetes 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.

Association analyses using the merged 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 9523 subjects (8842 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 minor allele frequencies were less than 0.05 and subjects whose call rates were less than 95% or identity by state was more than 0.8. The remaining 9521 subjects with 272,795 SNP genotypes were used for the analyses, and phenotypes of 7804 relatives were available.

In the Korean population, about 9.9% of adults over 30 years of age were expected to have T2D in 2009 [29], and the heritability of T2D has been reported to be approximately 26% [30]. Therefore, we set the prevalence and heritability values at 0.099 and 0.26, respectively, and calculated CEs for the 9521 subjects using the T2D status of their relatives. Based on these CEs, we selected 1000 cases and 4000 controls with S1 and S3. To adjust for population substructure, we calculated a genetic relationship matrix and applied the EIGENSTRAT approach [31]. We obtained the top 10 principal component scores with the largest eigenvalues, and they were included as covariates. We also included sex, age, and squared age as covariates.

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 monotonically increasing according 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. Six strategies, S1 to S6, which we described in the Materials and Methods 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 2000 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 I). Figures 4 and 5 and Tables II and III 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 IV (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 V (scenario 2) and Table VI (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 IV); that is, S3 and S6 show no significant difference in performance when pedigrees are composed of only nuclear families with the same structure. However, because 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 Tables V and VI, 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.

Robustness of conditional expectation 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 VII 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 performed a GWAS for T2D using the merged data to compare the performance between selection strategies that we considered in simulation study. The QQ plots in Figure 7 show that GWASs 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 [32–36], and we conducted the other GWASs with permuted phenotypes. Figure S1 in the supporting information shows QQ plots from GWASs with permuted phenotypes, and we can conclude that statistical testing is robust against such problems. Figure 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 GWASs using all subjects, followed by GWASs using cases and controls ascertained with S3. Table VIII 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 lead to substantial improvement of power for GWAS.

Discussion

Many studies have reported that family history of a disease is related to statistical power [32–35]. 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 [37]. 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 [38–42], 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 [20–25], 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 T2D. These approaches can be used with existing software such as MTG2 [43].

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 [44]. 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 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 [16], and the distribution of unknown liability scores can be chosen by comparing several candidate link functions based on the Akaike information criteria [45]. 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 [46]. Alternatively, we can consider the use of generalized estimating equations [47]. However, family histories of diseases have a highly unbalanced structure, which often leads to slow convergence or non-convergence of maximum likelihood estimations or to inflated statistical inferences [48]. 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 [36]. Fortunately, homogeneity tests between cases and controls are statistically valid as long as the estimates of odds ratio are carefully interpreted [36].

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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF), as funded by the Korean Government (NRF-2014S1A2A2028559) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2016R1D1A1A09919610).

References

1. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*. 2008; 24:133–141. [PubMed: 18262675]
2. Metzker ML. Sequencing technologies—the next generation. *Nature Reviews Genetics*. 2010; 11:31–46.
3. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biology*. 2011; 12:1.
4. Moore GE. Cramming more components onto integrated circuits. *Proceedings of the IEEE*. 1998;86.
5. Maher B. The case of the missing heritability. *Nature*. 2008; 456:18–21. [PubMed: 18987709]
6. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A. Finding the missing heritability of complex diseases. *Nature*. 2009; 461:747–753. [PubMed: 19812666]
7. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*. 2001; 69:124–137. [PubMed: 11404818]
8. Antoniou AC, Easton DF. Polygenic inheritance of breast cancer: implications for design of association studies. *Genetic Epidemiology*. 2003; 25:190–202. [PubMed: 14557987]
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; 29:51–67. [PubMed: 15892093]
10. Li M, Boehnke M, Abecasis GR. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *The American Journal of Human Genetics*. 2006; 78:778–792. [PubMed: 16642434]
11. Risch N. Implications of multilocus inheritance for gene–disease association studies. *Theoretical Population Biology*. 2001; 60:215–220. [PubMed: 11855955]
12. Edwards J. Familial predisposition in man. *British Medical Bulletin*. 1969; 25:58–64. [PubMed: 5782760]
13. Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*. 1965; 29:51–76.
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; 88:294–305. [PubMed: 21376301]
16. Bliss CI. The method of probits. *Science*. 1934; 79:38–39. [PubMed: 17813446]

17. Wilhelm S, Manjunath B. tmvtnorm: truncated multivariate normal and student t distribution. R Package Version. 2013; 1:4–8.
18. Vazquez A, Bates D, Rosa G, Gianola D, Weigel K. Technical note: an R package for fitting generalized linear mixed models in animal breeding. *Journal of Animal Science*. 2010; 88:497–504. [PubMed: 19820058]
19. Sargolzaei M, Iwaisaki H. Comparison of four direct algorithms for computing inbreeding coefficients. *Animal Science Journal*. 2005; 76:401–406.
20. Guey LT, Kravic J, Melander O, Burt 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; 35:236–246. [PubMed: 21308769]
21. 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; 35:790–799. [PubMed: 21922541]
22. Barnett IJ, Lee S, Lin X. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic Epidemiology*. 2013; 37:142–151. [PubMed: 23184518]
23. Nebert DW. Extreme discordant phenotype methodology: an intuitive approach to clinical pharmacogenetics. *European Journal of Pharmacology*. 2000; 410:107–120. [PubMed: 11134663]
24. 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; 95:1605–1610.
25. Risch NJ, Zhang H. Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *American Journal of Human Genetics*. 1996; 58:836. [PubMed: 8644748]
26. Gelman, A., Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press; 2006.
27. Cochran WG. Some methods for strengthening the common χ^2 tests. *Biometrics*. 1954; 10:417–451.
28. Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics*. 1955; 11:375–386.
29. Kim DJ. The epidemiology of diabetes in Korea. *Diabetes & Metabolism Journal*. 2011; 35:303–308. [PubMed: 21977448]
30. 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; 42:139–145. [PubMed: 10064092]
31. 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; 38:904–909. [PubMed: 16862161]
32. Antoniou AC, Easton DF. Polygenic inheritance of breast cancer: implications for design of association studies. *Genetic Epidemiology*. 2003; 25:190–202. [PubMed: 14557987]
33. 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; 29:51–67. [PubMed: 15892093]
34. 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; 78:778–792. [PubMed: 16642434]
35. Risch N. Implications of multilocus inheritance for gene-disease association studies. *Theoretical Population Biology*. 2001; 60:215–220. [PubMed: 11855955]
36. Bjørnland, T. Statistical methods for genetic association studies under the extreme phenotype sampling design. <https://arxiv.org/abs/1701.01286>
37. 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; 8:1273–1288. [PubMed: 9872982]
38. Wang J, Shete S. Analysis of secondary phenotype involving the interactive effect of the secondary phenotype and genetic variants on the primary disease. *Annals of Human Genetics*. 2012; 76:484–499. [PubMed: 22881407]

39. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin M-R, Coin LJ. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One*. 2012; 7:e34861. [PubMed: 22567092]
40. van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype–phenotype analysis for genome-wide association studies. *PLoS Genetics*. 2013; 9:e1003235. [PubMed: 23359524]
41. Li H, Gail MH. Efficient adaptively weighted analysis of secondary phenotypes in case–control genome-wide association studies. *Human Heredity*. 2012; 73:159–173. [PubMed: 22710642]
42. Schifano ED, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics*. 2013; 92:744–759. [PubMed: 23643383]
43. Lee SH, Van Der Werf J. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics*. 2016; 32:1420–1422. [PubMed: 26755623]
44. Benckek PH, Morris NJ. How meaningful are heritability estimates of liability? *Human Genetics*. 2013; 132:1351–1360. [PubMed: 23867980]
45. Bozdogan H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*. 1987; 52:345–370.
46. McCulloch, CE., Neuhaus, JM. *Generalized Linear Mixed Models*. Wiley Online Library; 2001.
47. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
48. Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G. *Longitudinal Data Analysis*. CRC Press; 2008.

Appendix A

A.1. Calculation of the conditional expectation

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

$$E(L_i | \mathbf{I}_{A_i}(\mathbf{L}_i) = \mathbf{1}_{n_i+1}).$$

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

$$f(\mathbf{L}_i) = |2\pi \sum_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{L}_i' \sum_i^{-1} \mathbf{L}_i\right), \text{ where } \sum_i = \text{cov}(\mathbf{L}_i) = h^2 \Psi_i + (1-h^2) \mathbf{I}_{n_i}.$$

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

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

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

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

where $\mathbf{t}_i = (t_{i1}^r \cdots t_{in_i}^r, t_i)^r$. We let $\xi_i = \Sigma_i \mathbf{t}_i$ and then the exponential term of MGF can be simplified to

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

and MGF becomes

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

We let σ_{ijk} indicate the (j, k) th element of Σ_i and $F_{ik}(x)$ indicate a marginal PDF for the k th element of $\mathbf{L}_i$ of the conditional PDF, $f_{\alpha_i}(\mathbf{L}_i)$; that is,

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

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

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

then the CE for subject i can be calculated by

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

A.2. Derivation of $F_{ik}(x)$

The $(n_i + 1)$ -dimensional liability vector, $\mathbf{L}_i$, can be partitioned into $(\mathbf{L}_i)_{-j}$ and L_{ij}^r for $j = 1, \dots, n_i$ 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:

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

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

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

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

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

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

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

Therefore, the multivariate marginal PDF of L_i becomes

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

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

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

Figure 2. Disease allele frequencies (DAFs) according to conditional expectations (CEs). (A, B) DAFs for cases and controls, respectively. All subjects were sorted with CE and classified to five different groups with CEs.

Figure 3. Quantile–quantile (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 2000 replicates.

Figure 4. Quantile–quantile (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 2000 replicates.

Figure 5.

Quantile–quantile (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 2000 replicates.

Figure 6.

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

Figure 7. Quantile–quantile plots for the results from genome-wide association study (GWAS) of type 2 diabetes. [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 8. Manhattan plots for the results from genome-wide association study (GWAS) of type 2 diabetes. [Colour figure can be viewed at wileyonlinelibrary.com]

Table 1

Empirical type I error estimates for scenario 1.

h^2	q	Significance levels	S1 ¹	S2 ²	S3 ³	S4 ⁴	S5 ⁵	S6 ⁶
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

Scenario 1 was considered for family structures of subjects' relatives. The empirical type I errors were estimated with 2000 replicates, and heritabilities were set to be 0.2 and 0.4.

¹S1: Cases and controls were randomly selected from affected and unaffected subjects, respectively.

²S2: Affected subjects with the highest conditional expectations (CEs) were selected as cases, and controls were randomly selected.

³S3: Affected (unaffected) subjects with the highest (lowest) CEs were selected as cases (controls).

⁴S4: Cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls.

⁵S5: Affected (unaffected) subjects with the lowest (highest) CEs were selected as cases (controls).

⁶S6: Affected (unaffected) subjects with the largest (smallest) number of affected first-degree relatives were selected as cases (controls).

Table II

Empirical type I error estimates for scenario 2.

h^2	q	Significance levels	S1 ¹	S2 ²	S3 ³	S4 ⁴	S5 ⁵	S6 ⁶
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

Scenario 2 was considered for family structures of subjects' relatives. The empirical type I errors were estimated with 2000 replicates, and heritabilities were set to be 0.2 and 0.4.

¹ S1: Cases and controls were randomly selected from affected and unaffected subjects, respectively.

² S2: Affected subjects with the highest conditional expectations (CEs) were selected as cases, and controls were randomly selected.

³ S3: Affected (unaffected) subjects with the highest (lowest) CEs were selected as cases (controls).

⁴ S4: Cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls.

⁵ S5: Affected (unaffected) subjects with the lowest (highest) CEs were selected as cases (controls).

⁶ S6: Affected (unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases (controls).

Table III

Empirical type I error estimates for scenario 3.

h^2	q	Significance levels	S1 ¹	S2 ²	S3 ³	S4 ⁴	S5 ⁵	S6 ⁶
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

Scenario 3 was considered for family structures of subjects' relatives. The empirical type I errors were estimated with 2000 replicates, and heritabilities were set to be 0.2 and 0.4.

¹ S1: Cases and controls were randomly selected from affected and unaffected subjects, respectively.

² S2: Affected subjects with the highest conditional expectations (CEs) were selected as cases, and controls were randomly selected.

³ S3: Affected (unaffected) subjects with the highest (lowest) CEs were selected as cases (controls).

⁴ S4: Cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls.

⁵ S5: Affected (unaffected) subjects with the lowest (highest) CEs were selected as cases (controls).

⁶ S6: Affected (unaffected) subjects with the largest (smallest) number of affected first-degree relatives were selected as cases (controls).

Empirical power estimates for scenario 1.

Table IV

h^2	q	Significance levels	S1 ¹	S2 ²	S3 ³	S4 ⁴	S5 ⁵	S6 ⁶
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

The empirical power levels were estimated with 2000 replicates at different levels of significance. We assumed that $h_a^2=0.005$, $h^2 = 0.2$ and 0.4 , and $q = 0.1$ and 0.2 . Data in bold are the best empirical power for each heritability, prevalence and significance level.

- ¹S1: Cases and controls were randomly selected from affected and unaffected subjects, respectively.
- ²S2: Affected subjects with the highest conditional expectations (CEs) were selected as cases, and controls were randomly selected.
- ³S3: Affected (unaffected) subjects with the highest (lowest) CEs were selected as cases (controls).
- ⁴S4: Cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls.
- ⁵S5: Affected (unaffected) subjects with the lowest (highest) CEs were selected as cases (controls).
- ⁶S6: Affected (unaffected) subjects with the largest (smallest) number of affected first-degree relatives were selected as cases (controls).

Empirical power estimates for scenario 2.

Table V

h^2	q	Significance levels	S1 ¹	S2 ²	S3 ³	S4 ⁴	S5 ⁵	S6 ⁶
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

The empirical power levels were estimated with 2000 replicates at different levels of significance. We assumed that $h_a^2=0.005$, $h^2 = 0.2$ and 0.4 , and $q = 0.1$ and 0.2 . Data in bold are the best empirical power for each heritability, prevalence and significance level.

- ¹ S1: Cases and controls were randomly selected from affected and unaffected subjects, respectively.
- ² S2: Affected subjects with the highest conditional expectations (CEs) were selected as cases, and controls were randomly selected.
- ³ S3: Affected (unaffected) subjects with the highest (lowest) CEs were selected as cases (controls).
- ⁴ S4: Cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls.
- ⁵ S5: Affected (unaffected) subjects with the lowest (highest) CEs were selected as cases (controls).
- ⁶ S6: Affected (unaffected) subjects with the largest (smallest) number of affected first-degree relatives were selected as cases (controls).

Empirical power estimates for scenario 3.

Table VI

h^2	q	Significance levels	S1 ¹	S2 ²	S3 ³	S4 ⁴	S5 ⁵	S6 ⁶
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

The empirical power levels were estimated with 2000 replicates at different levels of significance. We assumed that $h_a^2=0.005$, $h^2 = 0.2$ and 0.4 , and $q = 0.1$ and 0.2 . Data in bold are the best empirical power for each heritability, prevalence and significance level.

- ¹ S1: Cases and controls were randomly selected from affected and unaffected subjects, respectively.
- ² S2: Affected subjects with the highest conditional expectations (CEs) were selected as cases, and controls were randomly selected.
- ³ S3: Affected (unaffected) subjects with the highest (lowest) CEs were selected as cases (controls).
- ⁴ S4: Cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls.
- ⁵ S5: Affected (unaffected) subjects with the lowest (highest) CEs were selected as cases (controls).
- ⁶ S6: Affected (unaffected) subjects with the largest (smallest) number of affected first-degree relatives were selected as cases (controls).

Empirical relative power estimates for misspecified heritabilities and prevalences for scenario 3.

Table VII

h^2	q	Significance levels	S1 ¹	S2 ²	S3 ³	S4 ⁴	S5 ⁵
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

The empirical power levels were estimated with 2000 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 percentages. We assumed that $h_a^2=0.005$ and $(h^2, q) = (0.3, 0.3)$ for generating phenotypes. Four misspecified pairs of (h^2, q) were considered.

¹S1: Cases and controls were randomly selected from affected and unaffected subjects, respectively.

²S2: Affected subjects with the highest conditional expectations (CEs) were selected as cases, and controls were randomly selected.

³S3: Affected (unaffected) subjects with the highest (lowest) CEs were selected as cases (controls).

⁴S4: Cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls.

⁵S5: Affected (unaffected) subjects with the lowest (highest) CEs were selected as cases (controls).

Table VIII

Results from GWAS.

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

The significance level adjusted by Bonferroni correction is 1.872×10^{-7} and significant SNPs are indicated in bold type.

CHR, chromosome identifier; GWAS, genome-wide association study; POS, position; SNP, single-nucleotide polymorphism.