이 학 박 사 학 위 논 문

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

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

2019 년 2 월

서울대학교 대학원

협동과정 생물정보학과

김 원 지

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

by

Wonji Kim

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

Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University
Feb, 2019

Abstract

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

Wonji Kim
Interdisciplinary Program in Bioinformatics
The Graduate School
Seoul National University

Recent improvements in sequencing technology have enabled the investigation of so-called “missing heritability”, and a large number of affected subjects have been sequenced in order to detect significant associations between human diseases and genetic variants. However, the cost of genome sequencing is still high, and a statistically powerful strategy for selecting informative subjects would be useful.

Numerous methods for estimating heritability of dichotomous

phenotypes have been proposed. However, unlike quantitative phenotypes, heritability estimation for dichotomous phenotypes is computationally and statistically complex, and the use of heritability is infrequent. In particular, heritability estimates often suffer from substantial bias due to sampling scheme of family-based study. In family-based study, family members are often brought into a study via affected proband and therefore a proportion of affected relatives is larger than population prevalence. This bias refers to the ascertainment bias but there have been no much studies in adjusting method of ascertainment bias for heritability of dichotomous trait.

In this study, I propose a new statistical method for selecting cases and controls for sequencing studies based on disease family history in terms of improvement in statistical power of genetic association studies. I assume that disease status is determined by unobserved liability score. The liability threshold model assumes dichotomous phenotypes are determined by unobserved latent variables that are normally distributed, and our method consists of two steps: first, the conditional means of liability are estimated given the individual's disease status and those of their relatives with the liability threshold model, and second, the informative subjects are selected with the estimated conditional means. Our simulation studies showed that statistical power is substantially affected by the subject selection strategy chosen, and power is maximized when affected (unaffected) subjects with high (low) risks are selected as cases (controls). The

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

In addition, I developed a statistical method to estimate heritability of dichotomous phenotypes using a liability threshold model in the context of ascertained family-based samples. This model can be applied to general pedigree data. The proposed methods were applied to simulated data and Korean type-2 diabetes family-based samples, and the accuracy of estimates provided by the experimental methods was compared with that of established methods.

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

Student number: 2015-30118

Table of Contents

Chapter 1

Introduction

1.1 An Overview of Genetic Association Analyses of Dichotomous Phenotypes

Genetic association studies test association between a complex disease and genetic diversity in order to identify candidate causal genes or genomic regions [1]. At the level of a single nucleotide polymorphism (SNP), a higher frequency of certain alleles in a subject with a disease can be considered to mean that the SNP increases the risk of the disease. In addition to SNP, insertion/deletions (indels) and copy-number variants can be used as genetic variants for association studies and results can be interpreted in a similar way.

The Genome-wide association study (GWAS) was first proposed by Risch and Merikangas arguing that association studies are generally

more powerful than the linkage study in detecting genes of modest effect but requires much more markers to be tested [2]. They predicted that the complex diseases would require large-scale testing of association analysis. It also has been shown that genetic susceptibility to common complex disease includes many genes, most of which have small effects, leading to the importance of large-scale GWAS in a large-scale of sample sizes [3, 4]. Recently, several methods to improve statistical power of GWAS were proposed by accounting for sample structure in GWASs [5, 6]. They used linear mixed model and its extension to multi-loci was also developed [7].

As part of the effort for large-scale GWAS, several international projects have been undertaken. The international HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) genotyped for 3.1 million SNPs in DNA samples of 269 subjects from several populations which have ancestry of Africa, Asia and Europe [8]. It aims to develop a haplotype map of the human genome and figure out common patterns of human genetic variation involved in human disease. The 1000 Genomes Project (http://www.internationalgenome.org/) has validated approximately 84.4 million variants in 2,504 subjects from 26 populations consisting of African, American, Est Asian, European and South Asian [9]. It ran between 2008 and 2015, and aims to find most

genetic variants with frequencies of 1% or more in the studied populations. More recently, UK Biobank (https://www.ukbiobank.ac.uk/) was established and recruited 488,377 subjects aged between 40-69 years from across the United Kingdom [10]. DNA samples for 488,377 participants were genotyped at 807,411 variants containing SNPs and short indels. A web-based database, PheWeb (http://pheweb.sph.umich.edu:5000/), has provided thousands of GWAS results based on UK Biobank along with a fine display.

By April 2018, the GWAS has successfully discovered more than 69,000 SNP-trait associations (https://www.ebi.ac.uk/gwas/home/) [11-13]. These studies were rapidly growing in size and complexity, and in 5,152 studies, 3,378 publications were added to the GWAS catalog (Figure 1.1).

Figure 1.1 GWAS catalog as of 2018. All SNP-trait association with P-value $\leq 5\times10^{-8}$ were shown.

1.2 Heritability Estimation of Dichotomous Phenotypes

In 1950, Dempster and Lerner developed an algorithm to estimate the heritability of a binary trait [14], and their derivation was extended to the polychotomous traits by Gianola [15]. Their models were involved in the liability threshold model, which assumes that there is an underlying liability whose value is the sum of normally and independently distributed genetic and environmental components. In liability threshold models, the person is affected to the disease if his/her liability exceeds certain threshold of the underlying disease. A simulation study using Depmster's algorithm was performed by Van Vleck [16]. It was based on sib and parent-offspring family structure, and the estimated values of heritability were quite closed to the true values in a situation that a prevalence of a disease was ranged from 0.2 to 0.8 and the true heritability was below 0.7. There are several methods to estimate heritability of a dichotomous phenotype based on generalized linear mixed model (GLMM) such as logit-based algorithm [17, 18] and beta-binomial model [19, 20]. However, some of GLMM-based algorithm to estimate genetic variance components for multiple related relatives was developed but estimation of heritability is not

possible since environmental variance component is not included [21]. More recently, a method of estimating the proportion of phenotypic variance explained by a group of SNPs was proposed and it successively adjusted case-control ascertainment bias [22, 23].

1.3 The Purpose of This Study

The main purpose of this thesis is to develop statistical methods for genetic association analyses of dichotomous phenotype with related samples. In order to achieve this aim, I proposed two methods. One is a method to improve statistical power of GWAS by selecting informative cases and controls for DNA sequencing based on their family history. The other is intended to estimate heritability of a dichotomous phenotype based on liability threshold model for ascertained samples.

In the first study, I proposed a new statistical method for selecting informative cases and controls based on the disease status of their relatives. The proposed method is based on the conditional expectation of unobserved liability for subjects when the disease status of those subjects and their relatives are given. I assumed that the unobserved liability scores are normally distributed, and its conditional expectation will be the expectation of truncated normal distribution. In extensive simulation studies, I found that the statistical power is most increased when subjects with high and low risk are selected as cases and controls, respectively. Our methods were applied to GWAS of type-2 diabetes (T2D) and I compared the results for randomly selected samples and samples selected based on the proposed method.

In the second study, I proposed a method for heritability estimation of dichotomous phenotypes using liability threshold model. In particular, the proposed method can be applied to the ascertained samples by proband which refers to instances when family members are introduced to a study due to other family members already included in the study. Using the Expecteation-Maximization (EM) algorithm, the proposed method can estimate heritability and coefficients of covariates on the liability scale [14]. In addition, its statistical significance was assessed via a conditional expected score test (CEST) for the hypotheses if heritability is equal to zero or if coefficients of covariates are equal to zero. Using extensive simulation studies, I compared the proposed model to GCTA and I found that estimates of the proposed method are more generally unbiased for randomly selected families than that of GCTA. For ascertained samples, the proposed method works well similarly with that for randomly selected families, but GCTA produced substantial downward bias. I applied the proposed method to the T2D dataset to estimate the heritability of T2D in Korea population, and Lymphangioleiomyomatosis (LAM) dataset for GWAS.

1.4 Outline of the thesis

This thesis is organized as follows: Chapter 1 introduces to this study with an overview of GWAS and heritability estimation of dichotomous trait. Chapter 2 contains an example of GWAS for case-control study for LAM disease including a strategy for fine mapping. Chapter 3 is about a method to select informative subjects for DNA sequencing using family history to improve a statistical power. Chapter 4 deals with a method to estimate heritability of dichotomous phenotype for ascertained samples. Both Chapter 3 and 4 are based on the liability threshold model and population prevalence of a disease is required. Their performances were evaluated using extensive simulation study and applied to the real datasets. Finally, the summary and conclusions are presented in Chapter 5.

Chapter 2

Application of Genome-wide Association Study and Fine-mapping for Independent Samples

2.1 Introduction

Lymphangioleiomyomatosis (LAM) is a rare aggressive low-grade neoplasm which affects almost exclusively women at reproductive age or older and causes progressive cystic lung destruction leading to fatal respiratory failure in subjects with severe disease [24-29]. LAM is characterized by an abnormal proliferation of smooth muscle-like and epithelioid cells in innumerable tiny clusters in the lungs, in association with thin-walled cysts and lung parenchymal

destruction [30, 31]. Progressive cyst enlargement and inflammation contribute to decline in lung function measured as both decreased FEV_1 and DL_{CO}. The diagnosis of LAM is based on clinical features, chest computed tomography findings of thin-walled cysts, and either pathology seen on lung biopsy or elevated serum vascular endothelial growth factor D (VEGF-D) levels.

LAM occurs at high frequency (> 10%) in women with Tuberous Sclerosis Complex (TSC); and at much lower frequency in women (about 1 in 100,000) without that disorder, in which it is called sporadic (S-LAM). TSC is due to germline or somatic mutations in either *TSC1* (25%) or *TSC2* (75%) [32]. Tumor development in TSC follows the classic Knudson model of a germline mutation complemented by a somatic second hit mutation in the other corresponding allele in tumors [32, 33]. Limited data are available for S-LAM, but it appears that *TSC2* mutations are seen in the vast majority of S-LAM lesions. About 50% S-LAM subjects have kidney angiomyolipoma, a tumor which is seen in 70-80% of adults with TSC. Angiomyolipoma share histologic, expression, and genetic features with LAM, though are not identical pathologic lesions.

Genome-wide association studies (GWAS) are utilized to identify genetic variants and susceptibility loci associated with complex

traits and common diseases. Although there is no precedent for genetic influence on the development of S-LAM, I hypothesized that DNA sequence variants outside of *TSC2*/*TSC1* might be associated with disease risk, and go unrecognized due to the low prevalence of this disorder.

2.2 Materials and Methods

2.2.1 Discovery cohort

Over 600 female S-LAM patients were identified and collected through international solicitation from 2010 to 2014 from 14 countries (Table 2.1). S-LAM was diagnosed using standard diagnostic criteria [1-5, 7] by their treating physicians. Genomic DNA was extracted from saliva using the QIAamp DNA mini kit (Qiagen, Germany), and 479 S-LAM DNA samples were genotyped with the Infinium OmniExpress-24 v1.2 BeadChip, which assesses 716,503 SNPs across the entire genome. 34 non-white S-LAM subjects were excluded from further analyses.

Genotype data from the same genotyping chip were available for 1261 healthy female volunteers from the COPDGene Consortium, and were obtained from dbGaP (phs000951.v2.p2.c1). These COPDGene participants had smoked at least 10 pack years and were 45 to 80 years old, and were without known COPD [34, 35].

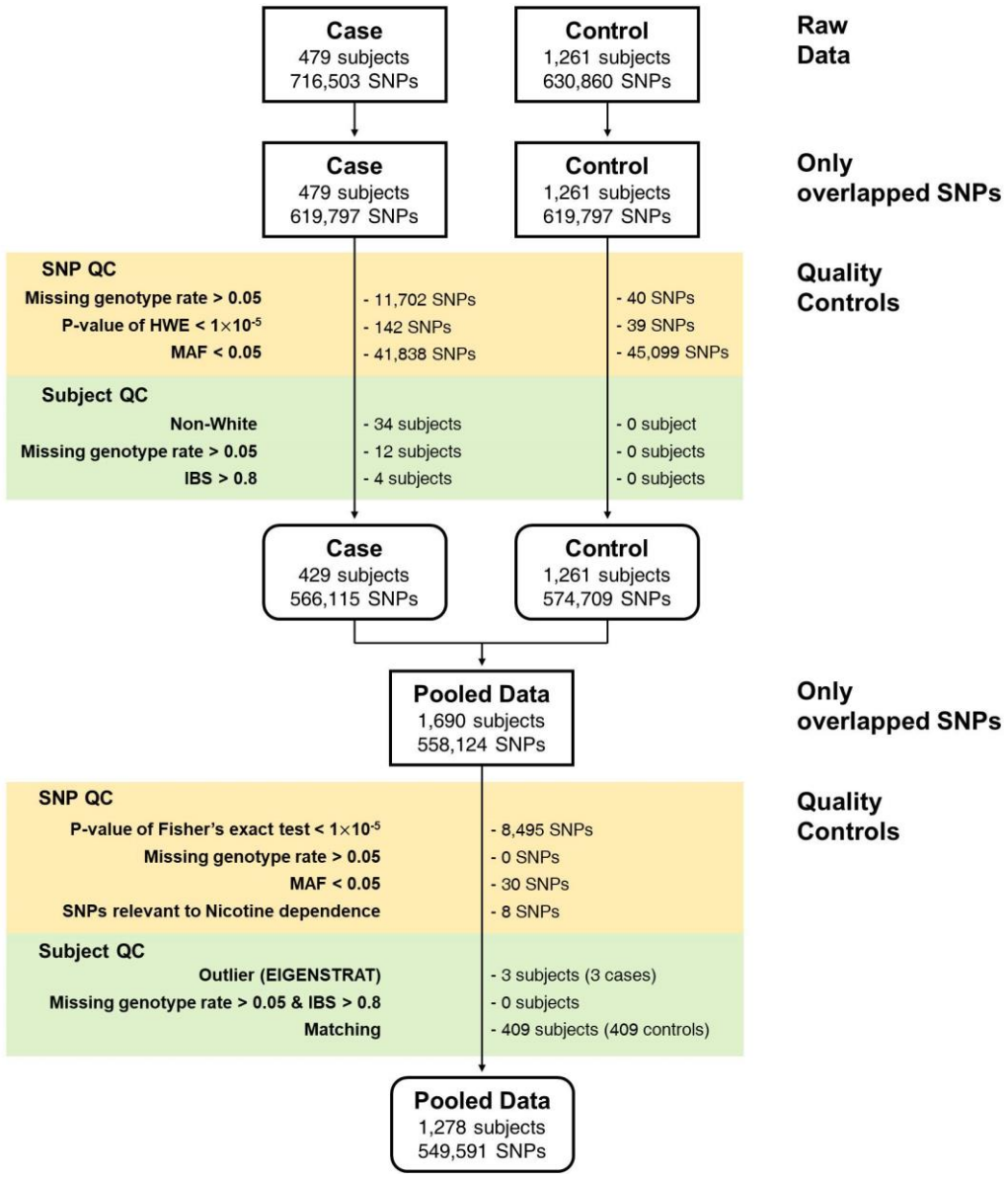
2.2.2 Quality control analyses of SNP genotype data

I evaluated the quality of SNPs and subjects in the discovery data set using PLINK [36] and ONETOOL [37]. I excluded all SNPs for which: the Hardy-Weinberg equilibrium (HWE) test [38] gave $P < 1\times10^{-5}$; minor allele frequency (MAF) was < 0.05; or genotype call rates were less than 95%. I also discarded any subjects whose missing genotype rates were > 5%, or showed identity-by-state > 80% with any other subject. These filtering procedures were first applied separately to cases and controls, and were repeated on the pooled dataset. In addition, any SNP showing a difference in missing data rate between cases and controls by Fisher's exact test [39], with $P < 1\times10^{-5}$ was removed. Last, EIGENSTRAT [40] was applied to the pooled data and principal component (PC) scores were calculated. PC scores were used to detect subjects with an outlying genetic background, and such outliers were then removed. These filters led to retention of 426 S-LAM cases and 852 female controls for analysis in the discovery phase with 549,599 SNP genotypes (Figure 2.1).

Table 2.1 Distribution of LAM patients according to their nationality

	Discovery LAM	Replication LAM
USA	190	196
France	54	0
Spain	40	0
Italy	35	0
United Kingdom	32	0
Germany	21	0
Australia	20	0
Poland	15	0
Israel	7	0
Canada	4	0
Panama	1	0
Puerto Rico	1	0
Scotland	1	0
Unknown	5	0
Total	426	196

Figure 2.1 Workflow of quality control for the LAM GWAS discovery data set. Multiple standard quality controls were performed for both cases (female S-LAM subjects) and controls (healthy women without COPD from COPDGene consortium) to exclude outlier SNPs and subjects.



2.2.3 Replication data

Replication analysis was done on an additional independent set of 196 non-Hispanic white (NHW) female S-LAM subjects, for the two SNPs identified in the discovery study, provided by one co-author (JM, Table 2.1). Careful scrutiny was performed by a third party to ensure that there was no overlap between the primary analysis population and the replication population. Genotyping was performed by TaqMan SNP genotyping assays C_832391_10 and C_27296040_10 for SNPs rs2006950 and rs4544201, respectively (ThermoFisher Scientific). Nine randomly selected S-LAM subjects from the discovery study were also genotyped by this method to confirm genotyping accuracy in the replication analysis. Their discovery study genotypes matched the TaqMan analysis genotypes perfectly, and these 9 subjects were not included in the replication analyses. 409 NHW healthy females from COPDGene Consortium who were not used for discovery analyses were used as controls for comparison in the replication study.

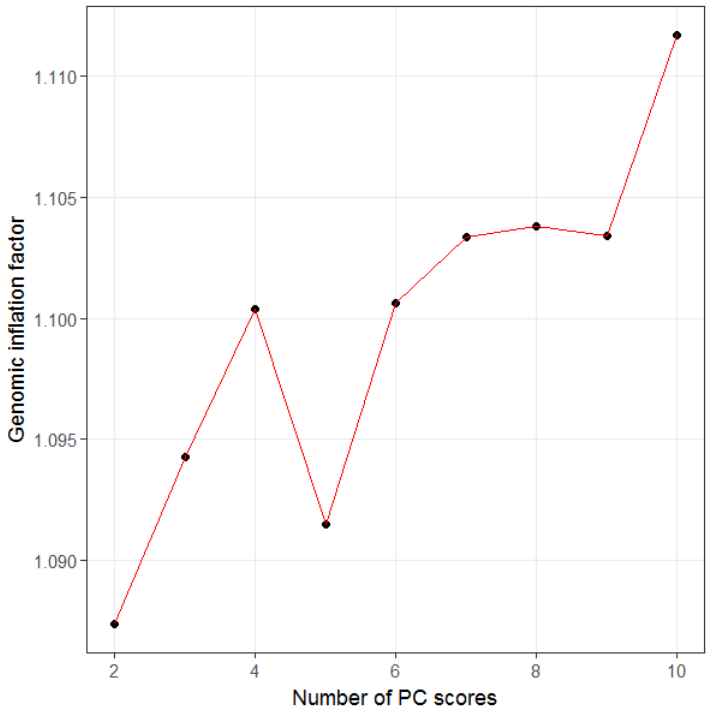
2.2.4 Statistical analyses with genetic data

GWAS analyses with discovery data were conducted using conditional logistic regression (CLR).

Principal Components (PC) Analysis scores were estimated with EIGENSTRAT [40], and used to adjust population substructure. CLR requires matching of cases and controls, and matching quality is affected by the number of PC scores matching. Each case was matched with two controls using the *Matching* R package [41]. Figure 2.2 shows that matching with age and two PC scores corresponding to the two greatest eigenvalues provide the variance inflation factor closest to 1. Thus CLR was conducted with cases and controls matched with age and 2 PC scores. CLR analyses were performed with the R package *survival* [42] and genome-wide significance was assessed by P-value < 5×10^{-8}.

I also conducted gene-based analyses to identify genes with significant association with S-LAM using the SKAT-O statistic [43]. SNPs within each gene were used to provide a SNP set file, and age, squared age and 10 PC scores were included as covariates.

Figure 2.2 Variance inflation factors according to the number of PC scores used for the discovery data. Cases and controls were matched with different numbers of PC scores (2 – 10 PC scores) and age, and CLR was applied to matched cases and controls. Variance inflation factors were calculated for different numbers of PC scores, and plotted against the numbers of PC scores.



2.2.5 Genotype imputation and statistical analyses with imputed genotypes

I imputed untyped SNPs located within 1 mega-base of the two genome-wide significant SNPs on chromosome 15 to do fine-mapping. Imputation was conducted using the Sanger Imputation Service (https://imputation.sanger.ac.uk). I used Haplotype Reference Consortium release v1.1 and considered predominantly European ancestry [44]. Pre-phasing and imputation was conducted with SHAPEIT [45] and the PBWT package [46], respectively, and imputation accuracy was evaluated with the INFO metric [47]. Imputed SNPs were filtered out if INFOs, MAFs or P-values for the HWE test were < 0.3, 0.05, or 1×10^{-5}, respectively. Linkage disequilibrium (LD) blocks were chosen by using Haploview with default options [48] and I applied CLR to all SNPs in the LD block with the genome-wide significant SNPs from the initial genotyping. Furthermore, I applied PICS software to imputed and genotyped SNPs within the 34kb LD block containing the genome-wide significant SNPs to calculate the probability of each individual SNP being the causal SNP [49].

2.2.6 Topologically associated domains (TADs) and chromatin interactions

To identify chromatin interactions in the region of interest on chromosome 15q26.2, I used a 3D genome browser (www.3dgenome.org) to predict TADs [50]. I checked for TADs around the genome-wide significant SNPs and protein coding genes belonging to each TAD were investigated. I analyzed TADs from four cell lines/tissues judged closest to LAM: (i) human fetal lung fibroblast (IMR90), (ii) lung-related tissues (LUNG), (iii) H1 derived mesenchymal stem cells (H1-MSC), and (iv) Human Umbilical Vein Endothelial Cells (HUVEC).

2.2.7 Statistical analyses with RNA sequencing data

Whole transcriptome RNA-Seq analysis was performed on one abdominal LAM tumor and four kidney angiomyoliopomas at the Broad Institute of Harvard and MIT. Briefly, mRNA-Seq was performed using polyA cDNA capture followed by cDNA library synthesis (Illumina Truseq RNA Library Prep Kit), and sequencing on Illumina machines, following the same methods and in the same facility in which the GTEx RNA-seq project occurred [24]. Read data was processed into FASTQ files with standard QC methods, and aligned to the genome (hg19, NCBI37) using Tophat v2.0.10 [51]. Fastq files were also converted into RSEM format [52]. RSEM values were compared to RNA-seq data from 2463 tumors of 27 different histologic types from the TCGA [53]. RPKM values for *NR2F2* were compared to the GTEx data set of normal human tissues (~7,000 samples from 53 normal tissue types, v6p release) [54].

2.2.8 Immunohistochemistry analyses

Immunochistochemistry was performed as described elsewhere [55] using a primary mouse monoclonal antibody against *NR2F2* [Abcam Cat.Num # ab41859 Concentration 1:100 (10ug/ml)]. Briefly, formalin-fixed, paraffin-embedded tumor sections were deparaffinized in xylene, rehydrated, and antigen retrieval was performed in EDTA (pH 8.0, Diagnostic BioSystems). Endogenous peroxidase activity was blocked with 3% H_2O_2, blocking was done with 5% goat serum, followed by incubation overnight with antibody at 4°C, washing in TBST, and incubation with anti-goat secondary antibody (Vector Labs, Burlingame, CA, dilution 1:300) The peroxidase reaction was developed using DAB substrate (DakoCytomation). Both LAM lung samples and kidney angiomyolipomas were stained by similar methods.

2.3 Results

2.3.1 GWAS analysis of S-LAM identifies two intergenic SNPs on chromosome 15

After multiple filtration steps and elimination of SNPs and samples as described in the Methods and shown in Figure 2.1, GWAS was performed on 426 S-LAM subjects and 852 control subjects from the COPDGene project, for 549,599 SNPs using CLR. Two non-coding SNPs rs4544201 and rs2006950 on chromosome 15 met genome-wide significance (rs4544201: P-value=8.51×10^{-10}; rs2006950: P-value=3.92×10^{-10}).

Quantile-quantile plots for CLRs and Manhattan plots demonstrated that the distribution of observed P-values met the expected distribution, with the exception of the two SNPs (Figure 2.3), indicating that the analyses were free of systematic P-value inflation. Multi-dimensional scaling plots indicated genetic similarity between cases and controls in the discovery analyses (Figure 2.4). Since the control COPDGene cohort were smokers, this association analysis might have been confounded by SNP alleles associated with nicotine addiction. I checked p-values for SNPs associated with nicotine addiction from the GWAS catalog [13] and other SNPs correlated with

those (r^2 >0.8) (Table 2.2). None of those SNPs showed a significant difference in allele frequency in the LAM and COPDGene cohorts, indicating that our findings are not confounded by nicotine addiction SNPs. Table 2.3 provides summaries for the two genome-wide significant SNPs.

rs4544201 and rs2006950 are located on 15q26.2, 11,563 nt apart, in an intergenic gene desert between *MCTP2* (1.1Mb away) and *NR2F2* (700kb away), that contains many lncRNAs (Figure 2.5). Both SNPs have minor and major alleles of A and G, and showed a lower minor allele frequency (MAF) in the S-LAM cohort than the control population. The odds ratios (ORs) of a single minor allele in the S-LAM cohort were 0.49 and 0.47 respectively, in comparison to the control population (Table 2.3). To adjust for the possible effect of the 'Winner's curse', I used br2 [56], and found that the bias-adjusted OR for rs4544201 and rs2006950 were 0.57 and 0.53, respectively.

Replication analysis was performed for the 2 SNPs with association with LAM using 196 additional non-Hispanic white (NHW) S-LAM patients and 409 NHW healthy females from COPDGene participants who were not used for discovery analyses. Similar ORs for association of the minor allele of these SNPs with S-LAM were seen in the replication data (Table 2.3, $OR_{rs4544201}$=0.33, $OR_{rs2006950}$ = 0.28).

Furthermore, I compared the MAFs of the 2 SNPs in LAM patients with those available from 7 other studies (composed of NHW European or USA populations), including the UKBiobank study of 337,199 individuals. The MAFs of the 2 SNPs in LAM patients were significantly smaller than those reported in every other cohort (Table 2.4).

rs4544201 and rs2006950 belong to the same LD block on 15q26.2 [48], and are strongly correlated (D'=0.977, r^2=0.854; Figure 2.6). To examine the potential association of other SNPs in this region with S-LAM, I used the genotyped SNP data to impute genotype data for all SNPs within 1 megabase of these two SNPs. Eighteen imputed SNPs in the 34kb LD block had P-values for association with LAM similar to rs4544201 and rs2006950 (Table 2.5).

To attempt to identify the causal SNP(s) among these SNPs with low P-values, I performed PICS analysis for all SNPs in Table 2.5, and the original two SNPs showing association. rs41374846 had both significant association with LAM, and the largest PICS probability (P_{PICS}=0.65, Table 2.6), making it the candidate causal SNP in this association [49].

Figure 2.3 Quantile-quantile plot and Manhattan plot for discovery LAM GWAS dataset. a) The observed distributions of P-values for 549,591 genotyped SNPs are plotted relative to the expected (null) distribution for each of CLR analyses. b) Each dot represents the P-value of a single SNP, plotted on the genome scale at bottom. The Y-axis value is the negative logarithm of the P-value for association between each genotyped SNP and LAM. Two SNPs on 15q26.2 met genome-wide significance ($P < 5 \times 10^{-8}$) by CLR.

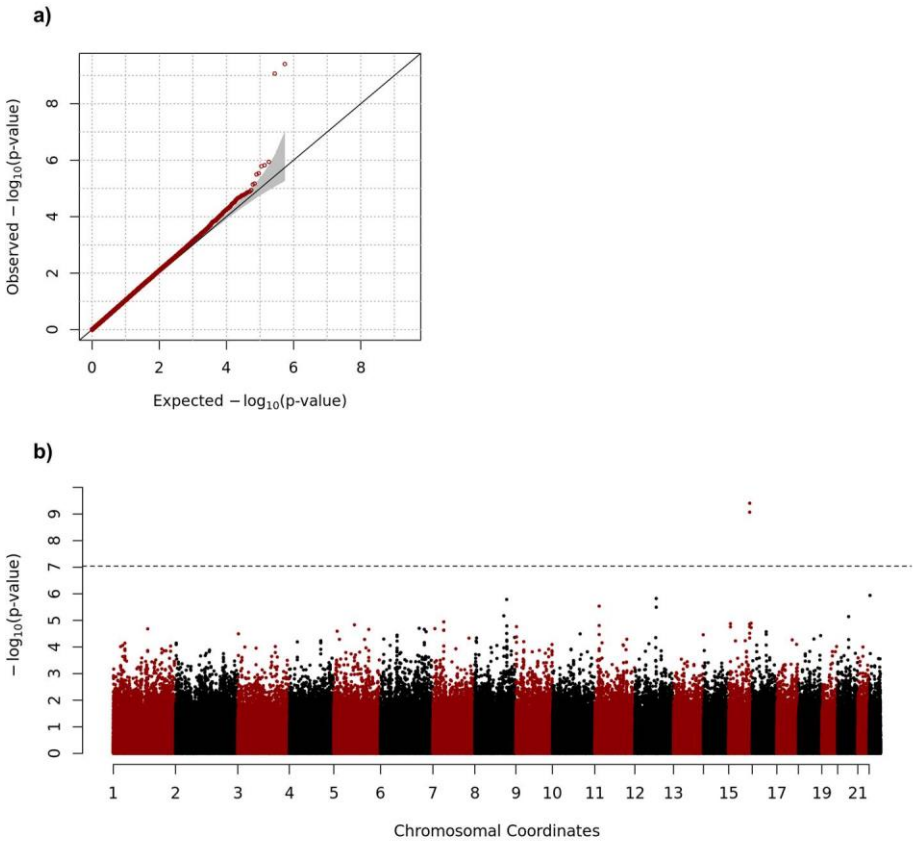


Figure 2.4 Multi-dimensional scaling plot. Multi-dimensional scaling plots were generated using a pool of our Discovery S-LAM cohort, our COPDGene controls, and 1000 Genome project data. Red and blue circles indicate S-LAM and COPDGene samples used for our discovery analyses, respectively, and grey circles represent participants for 1000Genome projects.

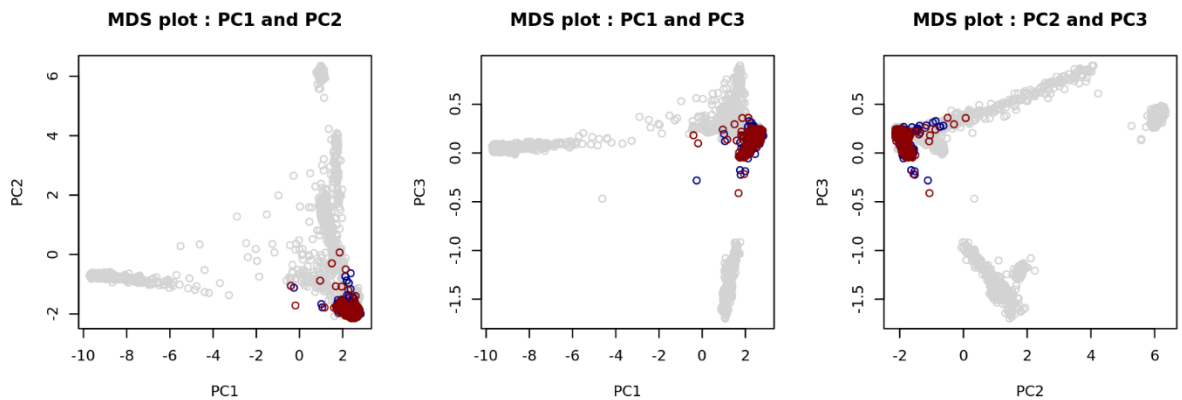


Table 2.2 P-values for SNPs associated with nicotine addiction. P values are shown in comparison of allele frequencies for the S-LAM discovery cohort and the COPDGene controls.

CHR	SNP	P-value
1	rs1060061	0.4885
6	rs9503551	0.0840
7	rs4285401	0.3263
8	rs804292	0.8145
8	rs6470120	0.1152
9	rs10491551	0.7217
4	rs10517300	0.6066
15	rs1051730	0.9759
21	rs2836823	0.1560

Figure 2.5 Genomic region on chromosome 15 containing the SNPs associated with LAM. a) Ideogram of chromosome 15. b) Three Mb region containing the SNPs associated with LAM. Manhattan plot at top shows P-values for SNPs in this region, including the two SNPs meeting genome-wide significance (red dots). There are 3 protein-coding genes *NR2F2*, *MCTP2*, and *SPATA8* which were represented in yellow shaded boxs, and many lncRNAs in this region. c) Expanded Manhattan plot of the 250kb region containing the genotyped and imputed SNPs showing association with LAM. SNP rs41374846 is indicated by purple, and other SNPs are colored according to their r^2 value in relation to rs41374846.

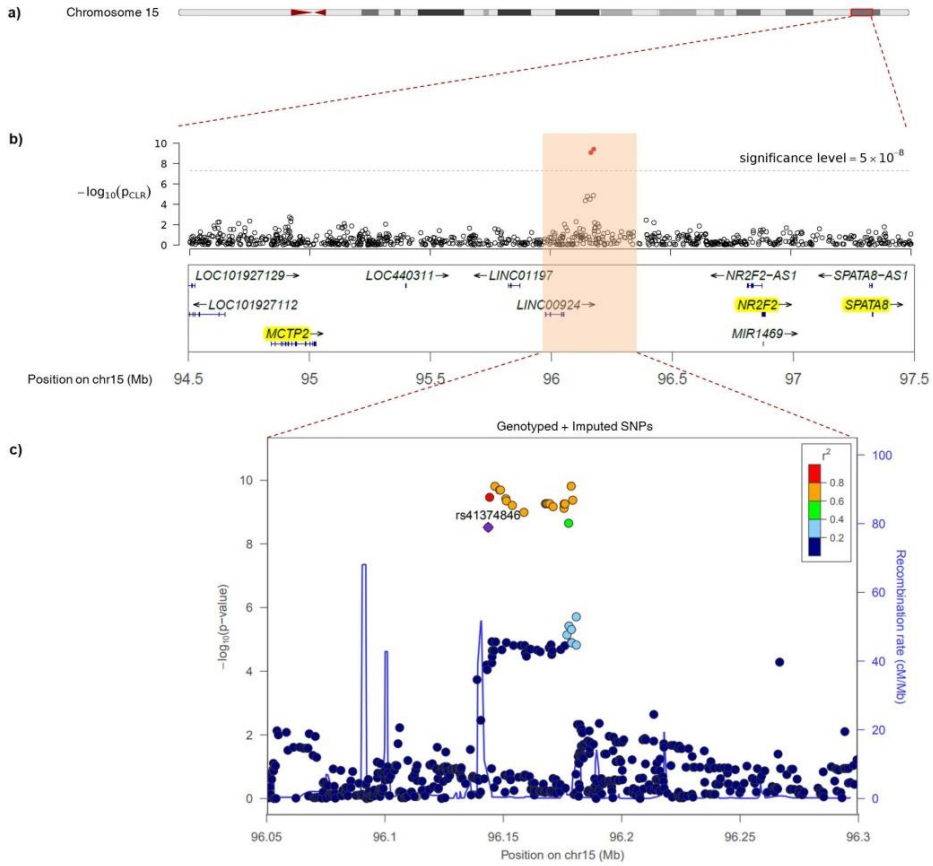

a)
Chromosome 15
b)
significance level = 5 × 10^-8
−log10(pCLR)
LOC101927129
LOC440311
LINC01197
NR2F2-AS1
SPATA8-AS1
LOC101927112
LINC00924
NR2F2
SPATA8
MCTP2
MIR1469
Position on chr15 (Mb)
94.5
95
95.5
96
96.5
97
97.5
c)
Genotyped + Imputed SNPs
rs41374846
r2
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
96.05
96.1
96.15
96.2
96.25
96.3
Position on chr15 (Mb)

Table 2.3 Genome-wide significant SNPs.

	rs4544201	**rs2006950**
Chromosome	15q26.2	15q26.2
SNP position (hg19)	96167827	96179390
Minor / Major alleles	A / G	A / G
Minor allele frequency		
S-LAM	0.1655	0.1420
Control	0.2750	0.2529
Genotype counts (AA / AG / GG / Missing)		
S-LAM	16 / 108 / 299 / 3	11 / 99 / 316 / 0
Control	62 / 343 / 444 / 3	58 / 315 / 479 / 0
Discovery data		
Odds ratio		
Original	0.4916	0.4732
Bias adjusted	0.5677	0.5315
P-value	8.51×10^{-10}	3.92×10^{-10}
Replication data		
Odds ratio	0.3288	0.2731
P-value	4.32×10^{-5}	1.56×10^{-5}

Table 2.4 Minor allele frequencies for SNPs rs4544201 and rs2006950 in multiple populations.

SNP	LAM patients			Normal		
	Data	N	MAF (95% CI)	Data	N	MAF (95% CI)
rs4544201	Discovery (USA/NHW/females)	190	0.1684 (0.131, 0.206)	COPDGene (USA/NHW/females)	1,258	0.2742 (0.257, 0.292)
	Discovery (EUR/NHW/females)	233	0.1631 (0.130, 0.197)	COPDGene (USA/NHW/males)	1,224	0.2774 (0.260, 0.295)
	Replication (USA/NHW/females)	186	0.1429 (0.107, 0178)	MESA-Lung* (USA/NHW/females)	1,153	0.2563 (0.238, 0.274)
				1000GP** (USA/NHW/females)	50	0.2600 (0.174, 0.346)
				1000GP** (EUR/NHW/females)	213	0.2300 (0.190, 0.270)
				ECLIPSE*** (EUR/NHW/females)	792	0.2563 (0.235, 0.278)
				UKBiobank† (EUR/NHW/both)	337,199	0.2605 (0.259, 0.262)

				GnomAD[‡] (EUR/NHW/both)	7,482	0.2601 (0.253, 0.267)
rs2006950	Discovery (USA/NHW/females)	190	0.1474 (0.112, 0.183)	COPDGene (USA/NHW/females)	1,261	0.2546 (0.238, 0.272)
	Discovery (EUR/NHW/females)	230	0.1377 (0.107, 0.169)	COPDGene (EUR/NHW/males)	1,226	0.2557 (0.238, 0.273)
	Replication (USA/NHW/females)	186	0.1148 (0.082, 0.147)	MESA-Lung[*] (USA/NHW/females)	1,128	0.2283 (0.211, 0.246)
				1000GP[**] (USA/NHW/females)	50	0.2300 (0.148, 0.312)
				1000GP[**] (EUR/NHW/females)	213	0.2160 (0.177, 0.255)
				ECLIPSE[***] (EUR/NHW/females)	792	0.2431 (0.222, 0.264)
				UKBiobank[†] (EUR/NHW/both)	337,199	0.2432 (0.242, 0.244)
				GnomAD[‡] (EUR/NHW/both)	7,496	0.2421 (0.235, 0.249)

* MESA = Multi-Ethnic Study of Atherosclerosis. Nonhispanic whites females were chosen and MAFs were calculated.

** 1000GP = 1000 Genome Project

*** ECLIPSE = Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points

† http://pheweb.sph.umich.edu:5000/

‡ http://gnomad.broadinstitute.org

Figure 2.6 Linkage disequilibrium (LD) block around the two genome wide significant SNPs, rs4544201 and rs2006950. Graph represents all genotyped SNPs in the 34kb LD block on chromosome 15q26.2. The color of each rectangle and number within indicates the level of LD between a pair of SNPs, with complete LD (D'=100%, no number shown) indicated by red, and no LD indicated by white.

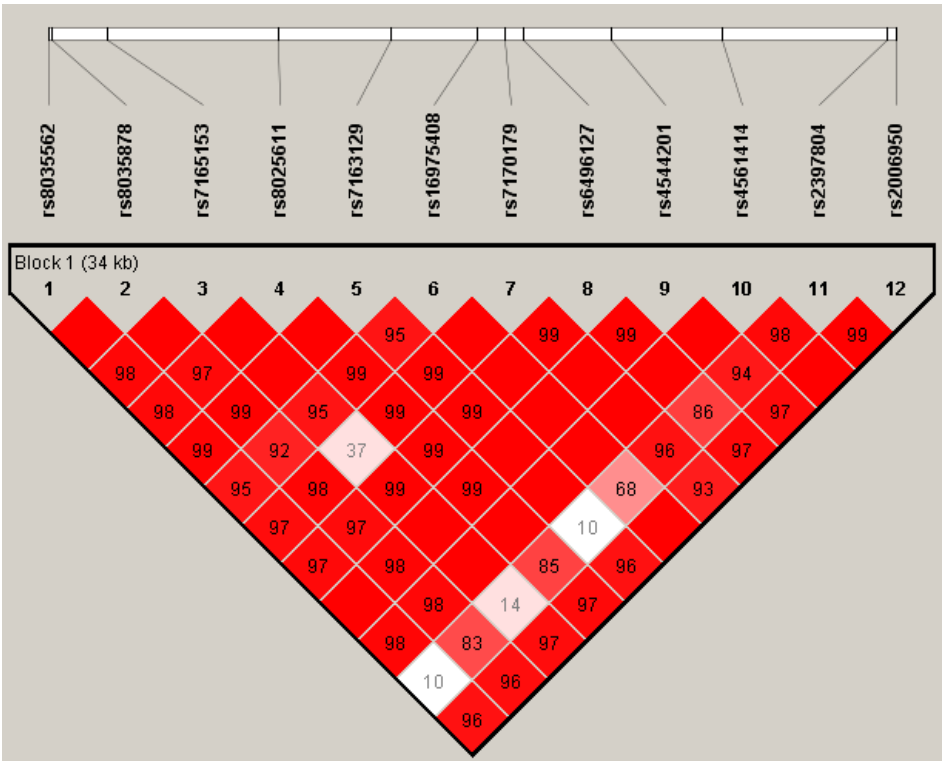


Table 2.5 Statistical analyses of imputed SNPs with CLR.

CHR	SNP	POS	Alleles*	MAF	INFO†	P-value for CLR‡
15	rs41374846	96143559	A/G	0.2605	0.9097	3.432×10^{-9}
15	rs59125351	96144157	G/T	0.2510	0.9771	3.229×10^{-10}
15	rs17581137	96146414	C/A	0.2336	0.9893	1.384×10^{-10}
15	rs6496126	96148439	C/G	0.2330	0.9890	1.814×10^{-10}
15	rs2397810	96148765	C/T	0.2330	0.9890	1.814×10^{-10}
15	rs10520790	96151040	T/G	0.2478	0.9958	3.571×10^{-10}
15	rs55804812	96151256	A/T	0.2475	0.9952	4.178×10^{-10}
15	rs16975389	96153782	C/T	0.2463	0.9967	5.801×10^{-10}
15	rs16975396	96158705	G/T	0.2466	0.9983	9.592×10^{-10}
15	rs4628911	96167905	T/C	0.2472	1.0000	5.147×10^{-10}
15	rs6496128	96168303	G/A	0.2472	1.0000	5.147×10^{-10}
15	rs8029996	96168770	A/G	0.2472	0.9998	5.147×10^{-10}
15	rs4551988	96169589	C/G	0.2472	0.9998	5.147×10^{-10}
15	rs58878263	96171069	A/C	0.2493	0.9979	6.361×10^{-10}
15	rs8040665	96175692	G/T	0.2487	0.9976	7.356×10^{-10}
15	15:96175733	96175733	A/G	0.2466	0.9975	5.224×10^{-10}
15	rs8040168	96176096	G/C	0.2466	0.9981	5.224×10^{-10}
15	rs17504029	96177670	T/A	0.2478	0.9876	1.900×10^{-10}

* Minor/Major alleles are listed.

† INFO is the metric about imputation quality provided by IMPUTE2.

‡ CLR was applied to imputed SNP genotype data to identify SNPs with significant association ($P < 5\times10^{-8}$) with S-LAM.

Table 2.6 PICS analysis to identify probable causal SNPs in the chr 15q region. SNP rs41374846 (shown in bold) was identified as the probable causal SNP, with the highest PICS probability. SNPs are sorted by PIC probability.

CHR	SNP*	POS	P-value	D'†	r^2‡	PICS probability
15	**rs41374846**	**96143559**	**3.432×10^{-9}**	1.0000	1.0000	0.6485
15	rs59125351	96144157	3.229×10^{-10}	0.9703	0.7941	0.0352
15	rs55804812	96151256	4.178×10^{-10}	0.9557	0.7758	0.0290
15	rs16975389	96153782	5.801×10^{-10}	0.9555	0.7700	0.0272
15	rs10520790	96151040	3.571×10^{-10}	0.9486	0.7698	0.0271
15	rs16975396	96158705	9.592×10^{-10}	0.9480	0.7581	0.0239
15	rs58878263	96171069	6.361×10^{-10}	0.9328	0.7287	0.0172
15	rs8029996	96168770	5.147×10^{-10}	0.9325	0.7230	0.0161
15	rs6496128	96168303	5.147×10^{-10}	0.9325	0.7230	0.0161
15	rs4628911	96167905	5.147×10^{-10}	0.9325	0.7230	0.0161
15	rs8040665	96175692	7.356×10^{-10}	0.9254	0.7171	0.0151
15	rs17581137	96146414	1.384×10^{-10}	0.9529	0.7125	0.0143
15	rs4544201	96167827	5.147×10^{-10}	0.9317	0.7116	0.0142
15	rs4551988	96169589	5.147×10^{-10}	0.9183	0.7113	0.0141
15	rs2397810	96148765	1.814×10^{-10}	0.9451	0.7008	0.0125
15	rs6496126	96148439	1.814×10^{-10}	0.9380	0.7005	0.0124
15	rs8040168	96176096	5.224×10^{-10}	0.9233	0.6887	0.0108

† $D' = D_{AB}/D_{\max}$ where D_{AB}: the frequency of the haplotype AB and $D_{\max}$: theoretical maximum difference between the observed and expected haplotype frequencies.

‡ r^2: squared correlation coefficient

2.3.2 Association of GWAS-significant SNPs with *NR2F2*

The majority of SNPs associated with human disease or other phenotypes are thought to cause the association through effects on enhancer or other regulatory element function of a coding gene within the topologically associated domain (TAD) containing the SNP [57]. To identify the TAD containing these SNPs, I used TAD information available for four tissues: IMR90 cells, a fetal lung myofibroblast cell line; lung tissue; H1-MSC, a mesenchymal stem cell line; and HUVEC, human umbilical vein endothelial cells (Figures 2.7-10). In all four of these cells/tissues, *NR2F2* was the only protein-coding gene within or near the boundary of the TAD containing the GWAS SNPs. This suggests that this SNP region may influence expression of *NR2F2* as its mechanism of association with S-LAM.

To examine this possibility in further detail, I conducted gene-based analyses of association of SNPs within all three protein-coding genes in the 2 MB region of chromosome 15 surrounding the GWAS-SNPs using SKAT-O. *NR2F2* was the only one of the three genes located in this chromosomal region that showed a significant association (P-value=0.03, Table 2.7).

NR2F2, also known as COUP-transcription factor II, encodes a member of the steroid/thyroid hormone superfamily of nuclear

receptors [58], and plays important roles in many developmental processes, including the neural crest [59], which is considered a potential candidate cell of origin of LAM [60], as well as in lymphangiogenesis and in angiogenesis [61]. Hence, I considered it a potential target of regulation by one of the SNPs showing a strong association with LAM (Table 2.6), and performed further studies.

Figure 2.7 Hi-C heatmap and TADs defined in IMR90 cells. The heatmap shows the degree of physical interaction defined by Hi-C analysis for genomic region pairs from a 3Mb region of chromosome 15q. A deeper red color at the intersection point reflects a greater degree of interaction between the two genomic regions. The dotted lines indicate probable TAD structures in this region. The two blue shaded regions at bottom indicate the genome wide significant SNP region (left) and *NR2F2* (right). The black circle reflects the interaction point between the SNP region and *NR2F2*.

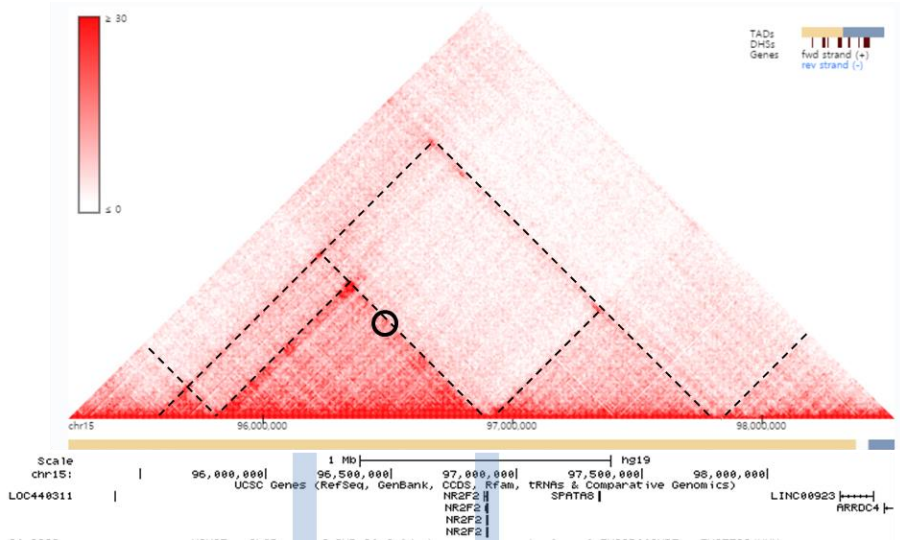


Figure 2.8 Hi-C heatmap and TADs defined in lung tissue. The heatmap shows the degree of physical interaction defined by Hi-C analysis for genomic region pairs from a 3Mb region of chromosome 15q. A deeper red color at the intersection point reflects a greater degree of interaction between the two genomic regions. The dotted lines indicate probable TAD structures in this region. The two blue shaded regions at bottom indicate the genome wide significant SNP region (left) and *NR2F2* (right). The black circle reflects the interaction point between the SNP region and *NR2F2*.

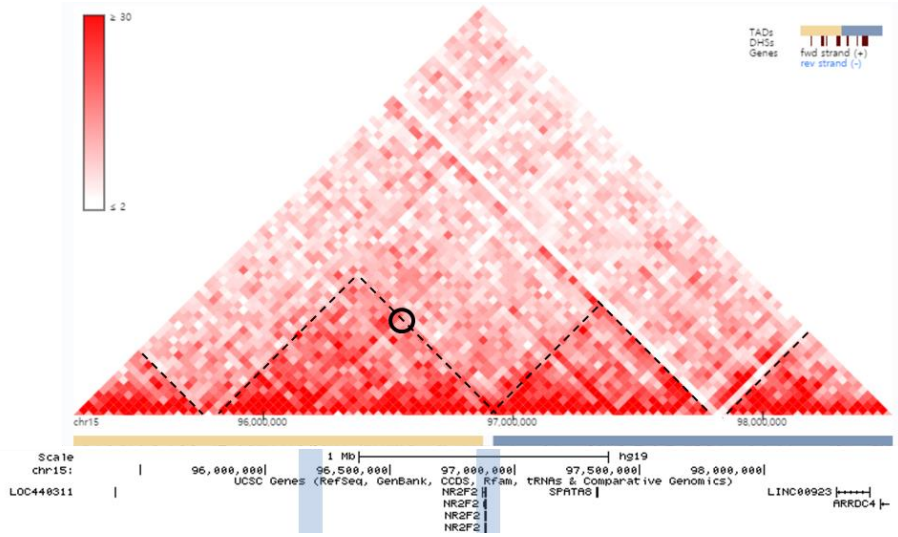


Figure 2.9 Hi-C heatmap and TADs defined in H1 derived mesenchymal stem cells (h1-MSC) cells. The heatmap shows the degree of physical interaction defined by Hi-C analysis for genomic region pairs from a 3Mb region of chromosome 15q. A deeper red color at the intersection point reflects a greater degree of interaction between the two genomic regions. The dotted lines indicate probable TAD structures in this region. The two blue shaded regions at bottom indicate the genome wide significant SNP region (left) and *NR2F2* (right). The black circle reflects the interaction point between the SNP region and *NR2F2*.

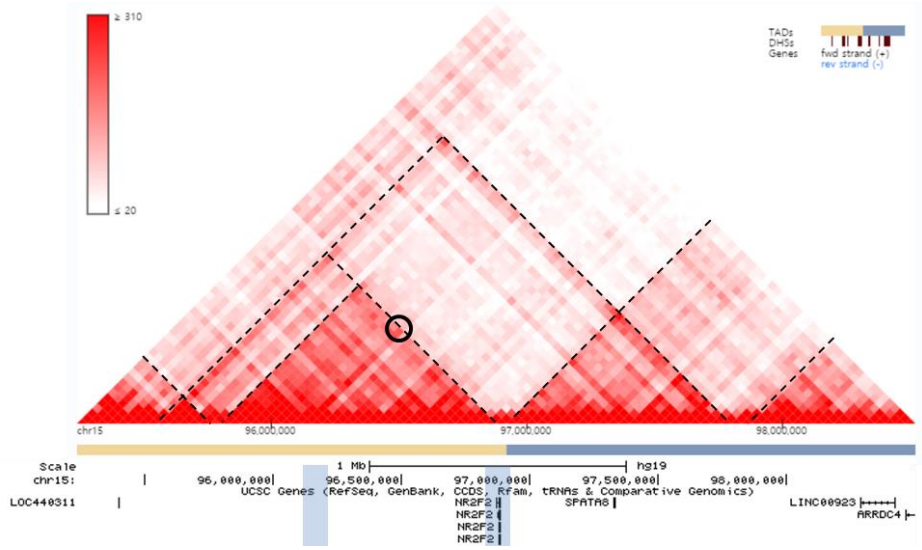


Figure 2.10 Hi-C heatmap and TADs defined in HUVEC cells. The heatmap shows the degree of physical interaction defined by Hi-C analysis for genomic region pairs from a 3Mb region of chromosome 15q. A deeper red color at the intersection point reflects a greater degree of interaction between the two genomic regions. The dotted lines indicate probable TAD structures in this region. The two blue shaded regions at bottom indicate the genome wide significant SNP region (left) and *NR2F2* (right). The black circle reflects the interaction point between the SNP region and *NR2F2*.

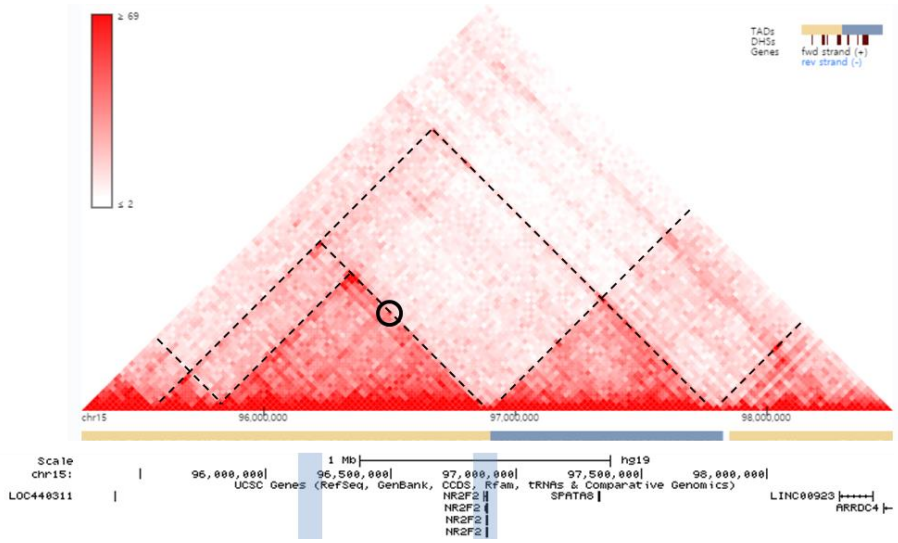


Table 2.7 Gene-based analyses of SNP association with LAM. Three protein-coding genes were found on chromosome 15 from 94.2 Mb to 98.2 Mb, the 2 Mb region surrounding the GWAS-SNPs, and gene-based analysis for association with LAM was performed using SKAT-O.

Gene	CHR	Start*	End†	Number of SNPs	P-value
NR2F2	15	96869157	96883492	5	0.0307
MCTP2	15	94774767	95027181	4	0.3579
SPATA8	15	97326619	97328845	3	0.5250

* Start position of the corresponding gene.

† End position of the corresponding gene.

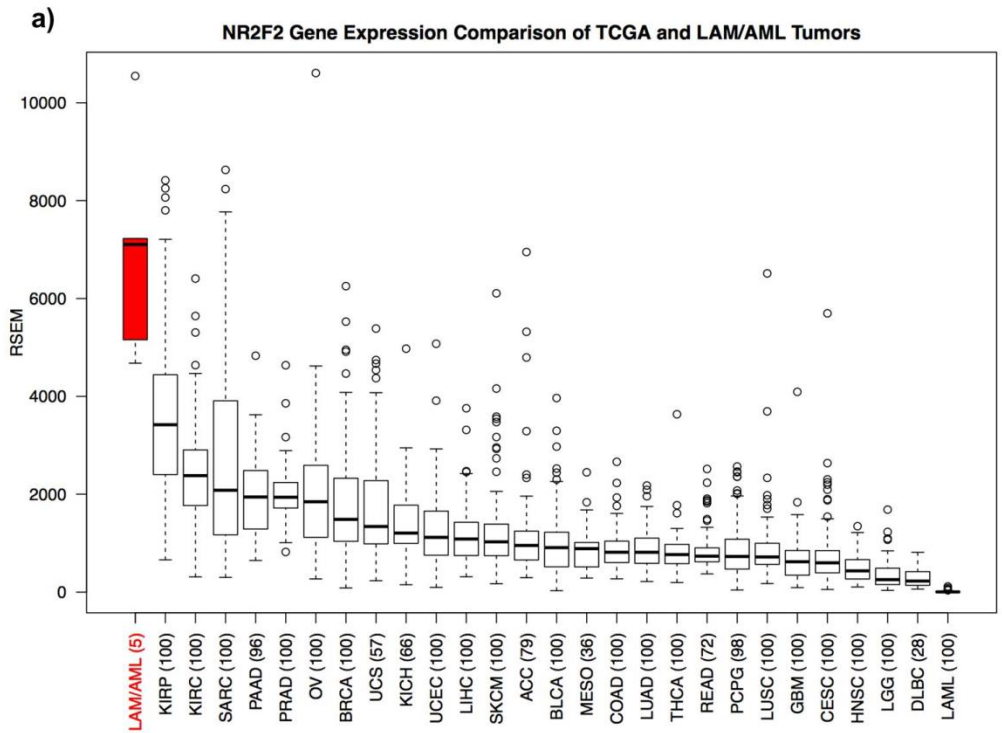
2.3.3 Analysis of *NR2F2* in kidney angiomyolipoma and LAM

Using RNA-seq data, I compared the gene expression of 4 four kidney angiomyolipomas and one abdominal LAM tumor with an extensive set of human cancers (from TCGA [53]), and normal tissues (from GTEX [54]) (Figure 2.11). *NR2F2* expression was higher in the LAM-related tumors than any TCGA cancer (Figure 2.11a), and was also relatively highly expressed in LAM-related tumors in comparison to normal tissues (Figure 2.11b, P-value=6.38×10^{-6}, Limma statistic) . In contrast, two other genes, *SPATA8* and *MCTP2*, that were next closest to the SNP region showing association with LAM (1.1 and 1.2Mb distant, Figure 2.4b) had no expression in the LAM-related tumors (data not shown).

Immunohistochemistry (IHC) analysis also demonstrated strong nuclear expression of *NR2F2* in both LAM lung and kidney angiomyolipoma sections (Figure 2.12).

Figure 2.11 Comparison of *NR2F2* expression in kidney angiomyolipoma/LAM with cancer (TCGA) and normal tissues (GTEx).

Boxplot figures are shown to compare expression of *NR2F2* in 4 angiomyolipoma and one abdominal LAM lesion with 2463 cancers of 27 types (from TCGA, brackets on x-axis include the number of samples analyzed per tumor type; abbreviations are explained in Table 2.8) in RSEM units (a); and with ~7,000 samples of 47 normal tissues (from GTEx) in RPKM units (b). Remarkably, NR2F2 gene expression is the highest compared to all TCGA tumors and higher compared to most GTEx normal tissues; similar to cervix, fallopian tubes, uterus and ovaries. The median value, interquartile range, and 95% ranges are shown, with outliers indicated by circles. In the X axis, the each number in brackets is the number of samples corresponding each tissue. Full terms for TCGA tumor abbreviations are explained in Table 2.8.


a)
NR2F2 Gene Expression Comparison of TCGA and LAM/AML Tumors
RSEM
10000
8000
6000
4000
2000
0
LAM/AML (5)
KIRP (100)
KIRC (100)
SARC (100)
PAAD (96)
PRAD (100)
OV (100)
BRCA (100)
UCS (57)
KICH (66)
UCEC (100)
LIHC (100)
SKCM (100)
ACC (79)
BLCA (100)
MESO (36)
COAD (100)
LUAD (100)
THCA (100)
READ (72)
PCPG (98)
LUSC (100)
GBM (100)
CESC (100)
HNSC (100)
LGG (100)
DLBC (28)
LAML (100)

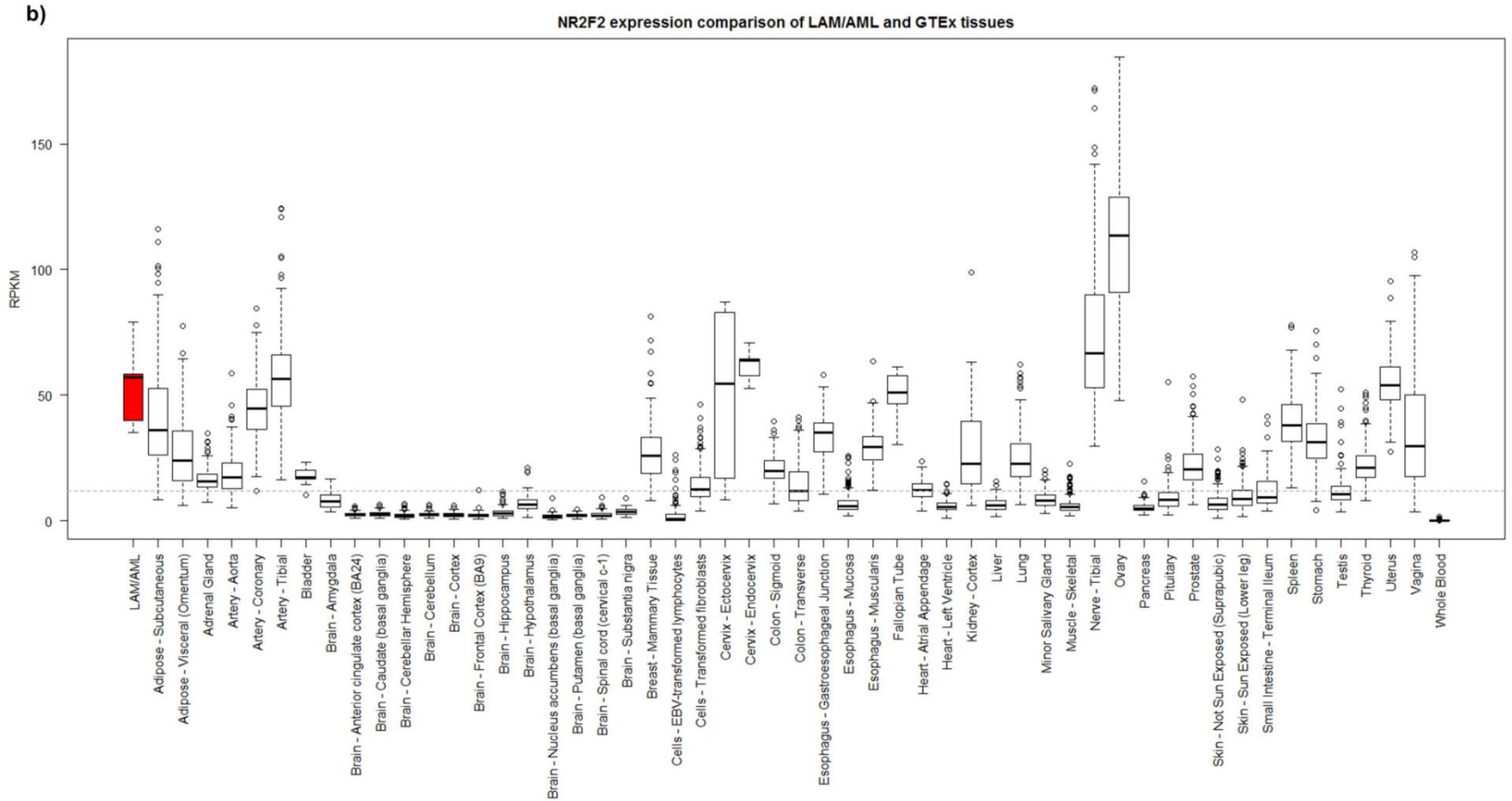
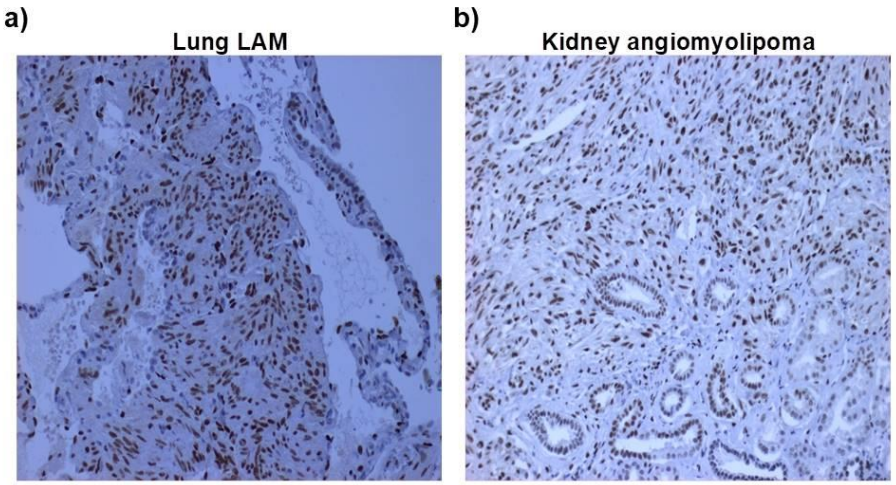

b)
NR2F2 expression comparison of LAM/AML and GTEx tissues
RPKM
0
50
100
150
LAM/AML
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cells - Transformed fibroblasts
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood

Table 2.8 TCGA tumor abbreviations

Abbreviation	Cancer type
KIRP	Kidney renal papillary cell carcinoma
KIRC	Kidney Renal Clear Cell Carcinoma
SARC	Sarcoma
PAAD	Pancreatic Adenocarcinoma
OV	Ovarian Serous Cystadenocarcinoma
BRCA	Breast Invasive Carcinoma
UCS	Uterine Carcinosarcoma
KICH	Kidney Chromophobe
UCEC	Uterine Corpus Endometrial Carcinoma
LIHC	Liver Hepatocellular Carcinoma
SKCM	Skin Cutaneous Melanoma
ACC	Adrenocortical Carcinoma
BLCA	Bladder Urothelial Carcinoma
MESO	Mesothelioma
COAD	Colon Adenocarcinoma
LUAD	Lung Adenocarcinoma
THCA	Thyroid Carcinoma
READ	Rectum Adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
LUSC	Lung Squamous Cell Carcinoma
GBM	Glioblastoma Multiforme
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
HNSC	Head and Neck Squamous Cell Carcinoma
LGG	Low Grade Glioma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
LAML	Acute Myeloid Leukemia

Figure 2.12 Immunohistochemistry for NR2F2 in LAM and kidney angiomyolipoma. Strong nuclear staining is seen in lung LAM cells (a) and angiomyolipoma cells (b) (brown stain). Some other cells also have nuclear staining for *NR2F2* but most do not.



2.4 Discussion

LAM occurs almost exclusively in women of childbearing age. Most LAM patients who come to medical attention are sporadic cases without TSC, and the origins of LAM in S-LAM patients are completely unknown. In the present study, I conducted a GWAS in a large cohort of S-LAM subjects. Two intergenic SNPs, rs4544201 and rs2006950, were identified in a 34kb LD block on chromosome 15, that met genome-wide significance for association with LAM (Table 2.3). The association was replicated in a validation population.

The SNPs with association to S-LAM lie in a gene desert on distal chromosome 15q26.2. The nearest protein-coding gene is *NR2F2*, 700kb away, and consideration of chromatin TADs in this region indicates that only *NR2F2* is in/on the border of the TAD region containing the SNPs showing association with S-LAM in four relevant cells/tissues, suggesting that these SNP alleles may influence *NR2F2* expression as the potential mechanism of their association with S-LAM development.

NR2F2 is an orphan nuclear receptor known to play important roles in both normal tissue development and in tumorigenesis [62], making it a promising candidate driver gene in LAM pathogenesis. LAM occurs nearly exclusively in women, and estrogen levels

influence LAM development and progression [63, 64]. siRNA knockdown of ERα (Estrogen Receptor) in MCF-7 breast cancer cells decreased *NR2F2* expression, while treatment with estradiol increased its expression [65]. This interaction between ERα and *NR2F2* may also play a role in LAM development.

NR2F2 is highly expressed in LAM and angiomyolipoma by RNA-Seq analysis in comparison to large cancer and normal tissue data sets, and *NR2F2* shows high expression with nuclear localization in both LAM and angiomyolipoma by IHC. Although I did not identify an eQTL relationship for any of the 20 SNPs associated with S-LAM for any gene in any normal tissue or cancer type [54], it is possible that such an eQTL relationship exists for LAM cells. I also note that the region of these SNPs contains several non-coding long RNAs, some antisense transcripts, and microRNA miR1469 (Figure 2.11a). It is possible that expression of one or more of these noncoding genes are affected by these SNP alleles, and have a role in LAM development, a possibility which requires further investigation.

Lymphatic involvement in LAM is a hallmark pathologic feature with LAM cell clusters in the lung showing marked enrichment for lymphatic vessels [66, 67]. VEGF-D is a probable driver of lymphatic vessel growth in LAM, as serum VEGF-D levels are

increased in the majority of LAM patients, and serves as a diagnostic biomarker of LAM [68]. In mice, *NR2F2* has been shown to be required, with *SOX18*, for the polarized expression of *PROX1* in a subset of endothelial cells within the cardinal vein at embryonic day 9.5, an event that leads to development of the lymphatic endothelium [69]. Hence there is also a potential connection between *NR2F2*, VEGF-D, lymphatic development, and LAM pathogenesis.

There are potential limitations to our study. Although our cohort of samples was large for a rare disease like S-LAM, it was of only moderate size for GWAS. Second, to collect sufficient LAM subjects, I employed a worldwide recruitment strategy for S-LAM patients of European origin. Although our controls were all from the USA, they were selected for European ancestry. In addition, I employed EIGENSTRAT to identify genetic outliers from both our S-LAM and control cohorts to further reduce genetic heterogeneity. Further functional analyses to confirm our hypothesis that *NR2F2* is the gene affected by this SNP is limited by the absense of a reliable LAM tumor cell line, the low abundance of LAM cells in LAM lung specimens, and lack of a LAM animal model.

In conclusion, our GWAS has identified non-coding SNPs on chr15q26.2 whose alleles are associated with S-LAM, that are located

in a TAD containing the orphan nuclear receptor *NR2F2*, suggesting a model in which these SNP alleles influence *NR2F2* expression and thereby LAM pathogenesis. *NR2F2* is relatively highly expressed in LAM and LAM-related tumors. *NR2F2* has not previously been implicated in LAM, and these novel and unexpected findings will hopefully lead to better understanding of the pathogenesis of this often progressive and lethal lung disorder.

This chapter was published in *Statistics in Medicine* as a partial fulfillment of Wonji Kim's Ph.D program.

Chapter 3

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

3.1 Introduction

Over the last several decades, DNA sequencing technologies have greatly improved, and the rate of decline in sequencing costs has even outpaced Moore's law [70-73]. 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 [74, 75]. 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 [76] assumes multiple rare disease susceptibility loci, suggesting that causal variants for each affected subject may be substantially different, and this genetic heterogeneity among affected subjects has also complicated genetic association analyses. Therefore, in spite of remarkable improvement in sequencing technology, development of efficient strategies for selecting informative subjects is still necessary, and various statistical methods have been investigated for use in genetic association studies.

Subjects with many affected relatives tend to contain more disease genotypes for heritable diseases, and it has been empirically shown that their ascertainment for genetic studies have often led to additional improvements in statistical power [77-80]. 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 [77-80]. 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 [77-80].

In this report, I propose a new statistical method for selecting informative subjects based on the disease status of their relatives [81] . 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 [82] or Falconer's liability threshold model [83]. 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 [84], and some literature shows their asymptotic relationship [23]. Heritability estimation at the observed data scale [84] is intuitively easier to understand, but its application to general family structures is not straightforward. Therefore, I 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 [85]. The unobserved liability scores are assumed to follow the normal distribution, and I calculate the conditional expectation with moment-based methods [86]. The proposed method can utilize the disease status of any type of relative, and using extensive simulation studies, I show that the statistical power is maximized when subjects with high and low risk are selected as cases and controls, respectively. The proposed methods were applied to genome-wide association studies (GWAS) for type-2 diabetes (T2D) with data collected from the Korea Association

REsource (KARE) project and Seoul National University Hospital in Korea (SNUH). The discovery of promising disease susceptibility loci illustrates the practical value of the proposed method.

3.2 Methods

3.2.1 Notations and the disease model

We assume that there are n independent subjects and that subject i has n_i relatives (i=1, …, n). I 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 HWE in a population. I 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 [83], 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 I consider the simple additive polygenic effect model. I 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 I 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, I can assume that

$$\mathbf{L}_i = \mathbf{Z}_i\beta_0 + \mathbf{G}_i\beta + \mathbf{P}_i + \mathbf{E}_i,$$

$$\mathbf{P}_i \sim MVN\left(\mathbf{0}_{n_i+1}, \sigma_g^2\mathbf{\Psi}_i\right), \mathbf{E}_i \sim MVN\left(\mathbf{0}_{n_i+1}, \sigma_e^2\mathbf{I}_{n_i+1}\right). \quad (1)$$

Here, $\mathbf{\Psi}_i$ indicates the kinship coefficient matrix for both subject *i* and his/her relatives. I denote the kinship coefficient between subject *i* and his/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/her relative j, respectively. The inbreeding coefficient, which ranges from 0 to 1, quantifies the departure from HWE and can be easily estimated using known pedigree by currently available R packages, e.g. *pedigreemm* [87, 88]. Then, $\boldsymbol{\Psi}_i^r$ and $\boldsymbol{\Psi}_i$ are defined by

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

and

$$\boldsymbol{\Psi}_i = \begin{pmatrix} 1+d_{i1}^r & \cdots & 2\pi_{i1n_i}^r & 2\pi_{i1} \\ \vdots & \ddots & \vdots & \vdots \\ 2\pi_{i1n_i}^r & \cdots & 1+d_{in_i}^r & 2\pi_{in_i} \\ 2\pi_{i1} & \cdots & 2\pi_{in_i} & 1+d_i \end{pmatrix}.$$

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/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

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

and that for the 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 I denote the cumulative function of a standard normal distribution by Φ and there are no covariate effects other than the intercept, I 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 I denote its variance by σ_z^2, c can be obtained by

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

3.2.2 Selection of samples with extreme phenotypes

Subjects with extreme phenotypes lead to improvement of statistical power in genetic association studies [89-93], and association analyses have often been conducted with such subjects. At the sample selection stage, genotypes of subjects are not known, and I assume β = 0 in equation (1). In particular, environmental factors can affect the dichotomous phenotypes and if their effects are known, I can then define the adjusted extreme phenotypes for dichotomous phenotypes by the following conditional expectation (CE) of liability scores:

$$\mathrm{CE} = E(L_i - Z_i\beta_0 | \mathbf{Y}_i, \mathbf{Z}_i)$$

CEs were calculated with a moment-based method [86] and the detailed algorithm is provided in the Appendix. Once I 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), **Z**, and β_0 are known. While h^2, q, and **Z** are often available *a priori*, the regression coefficients of environment effects are usually estimated from logistic regression, and they cannot be used as estimates of β_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 I assume that mean and variance for the cumulative normal distribution are 0 and 1.6, respectively, it is approximately equal to the logistic regression [94]. Therefore, if I 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.

3.2.3 Statistical power when the family history of disease is controlled

The statistical power for genetic association analysis with a case-control study design can be calculated when the relatives' phenotypes are conditioned. I consider the liability model in equation (1) and assume a major disease gene model. If I let q be the prevalence of the disease and I 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

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

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

and

$$P(G_{i'}|Y_{i'}=1,\mathbf{Y}_{i'}^r)=\sum_{\mathbf{G}_{i'}^r}P(G_{i'},\mathbf{G}_{i'}^r|Y_{i'}=1,\mathbf{Y}_{i'}^r)$$

$$=\sum_{\mathbf{G}_{i'}^r}\frac{P(Y_{i'}=1,\mathbf{Y}_{i'}^r|G_{i'},\mathbf{G}_{i'}^r)P(G_{i'},\mathbf{G}_{i'}^r)}{P(Y_{i'}=1,\mathbf{Y}_{i'}^r)}.$$

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

$$P(G_i|Y_i=1,\mathbf{Y}_i^r)$$

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

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

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

and

$$\mathrm{P}(G_{i'}=DD|Y_{i'}=1,\mathbf{Y}_{i'}^r)+0.5\mathrm{P}(G_{i'}=Dd|Y_{i'}=1,\mathbf{Y}_{i'}^r).$$

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

$$p_D^a=\frac{1}{n_a}\sum_{i=1}^{n_a}\{\mathrm{P}(G_i=DD|Y_i=1,\mathbf{Y}_i^r)+0.5\mathrm{P}(G_i=Dd|Y_i=1,\mathbf{Y}_i^r)\}$$

and

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

the statistical power for a Cochran Armitage test [95, 96] under the alternative hypothesis can be obtained from

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

If I 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

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

3.3 Simulation study

3.3.1 The simulation model

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 ($n \geq n_a + n_u$). I also assume that phenotypes for each subject's relatives are available, and I consider three different scenarios: (1) phenotypes of two parents and four siblings are known; (2) phenotypes of four grandparents, two parents, and four siblings are known; and (3) phenotypes of two parents and four siblings are known for half of the subjects, and phenotypes of four grandparents, two parents, and four siblings are known for the other half. Pedigrees for scenarios 1 and 2 are provided in Figure 3.1. The 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, I considered the disease model in equation (1). I 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 I let

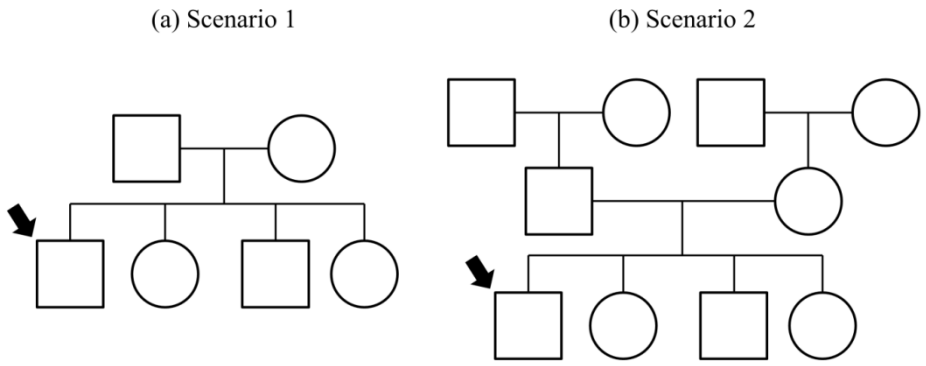
$$h^2 = \frac{2\beta^2 p_D p_d + \sigma_g^2}{2\beta^2 p_D p_d + \sigma_g^2 + \sigma_e^2} \text{ and } h_a^2 = \frac{2\beta^2 p_D p_d}{2\beta^2 p_D p_d + \sigma_g^2 + \sigma_e^2},$$

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, I 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 2,000 replicates at several significance levels. In each replicate, I 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, I additionally considered another strategy (S6), where the largest (smallest) number of affected first-degree relatives was selected as cases (controls). And then, I compared empirical sizes and powers using logistic regression.

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



3.3.2 Evaluation of selection strategy with simulated data

We investigated the effect of the selection strategy with simulated data. Six strategies, S1 to S6, which I described in the Method section, were used for genetic association analyses and were performed with the logistic regression. For each strategy, I selected 500 cases and 500 controls from 10,000 individuals, and empirical type-I errors and power were evaluated for each scenario with 2,000 replicates. Quantile-quantile (QQ) plots (Figure 3.2)
show that the nominal significance level was generally well preserved for scenario 1, and the empirical type-I error rates generally preserved the nominal significance level (Table 3.1). Figures 3.3-4 and Tables 3.2–3 show that the nominal significance levels were generally well preserved for scenarios 2 and 3 as well. Therefore, I 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. I assumed that h_a^2 = 0.005, h^2 = 0.2 or 0.4, and q = 0.1 or 0.2. Table 3.4 (scenario 1) shows that S3 was always the most efficient strategy among S1-S5, followed by S2 and S4. Interestingly, the statistical power estimates for S3 tended to be larger when the prevalence was larger and heritability was smaller, which indicates that

the proposed method would be useful for common diseases. S5 always gave the highest rates of false-negative findings, as this strategy minimizes differences in DAFs between cases and controls. Table 3.5 (scenario 2) and Table 3.6 (scenario 3) showed very similar patterns to scenario 1. Therefore, I concluded that cases and controls ascertained with S3 leads to substantial improvement in power.

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

Figure 3.2 Quantile-quantile (QQ) plots of simulated data for scenario 1. I 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

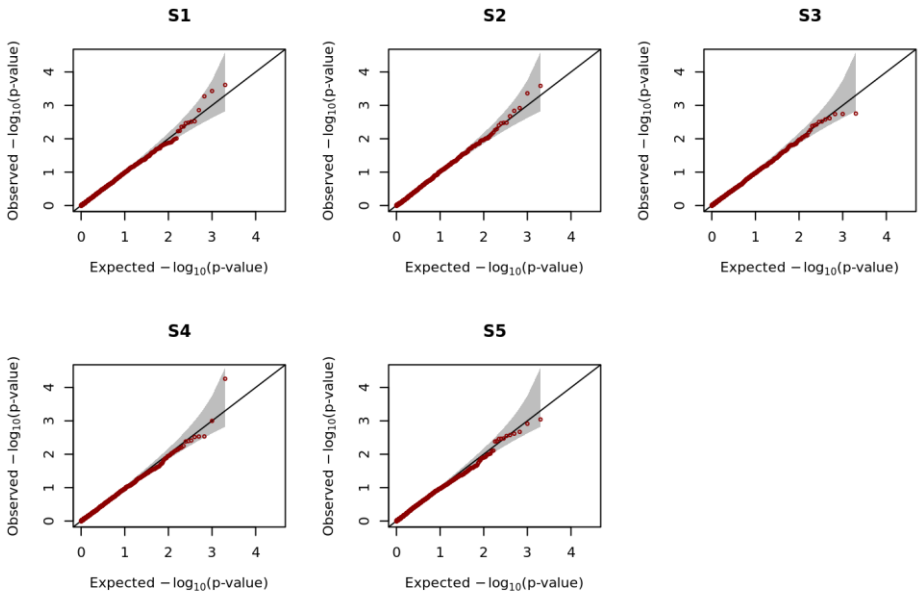


Table 3.1 Empirical type-I error estimates for scenario 1. Scenario 1 was considered for family structures of subjects' relatives. The empirical type-I errors were estimated with 2,000 replicates, and heritabilities were set to be 0.2 and 0.4.

h^2	q	Significance levels	S1[a]	S2[b]	S3[c]	S4[d]	S5[e]	S6[f]
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

[a] S1 : cases and controls were randomly selected from affected and unaffected subjects, respectively

[b] S2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected

[c] S3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)

[d] S4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls

[e] S5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)

[f] S6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls)

Figure 3.3 Quantile-quantile (QQ) plots of simulated data for scenario 2. I 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

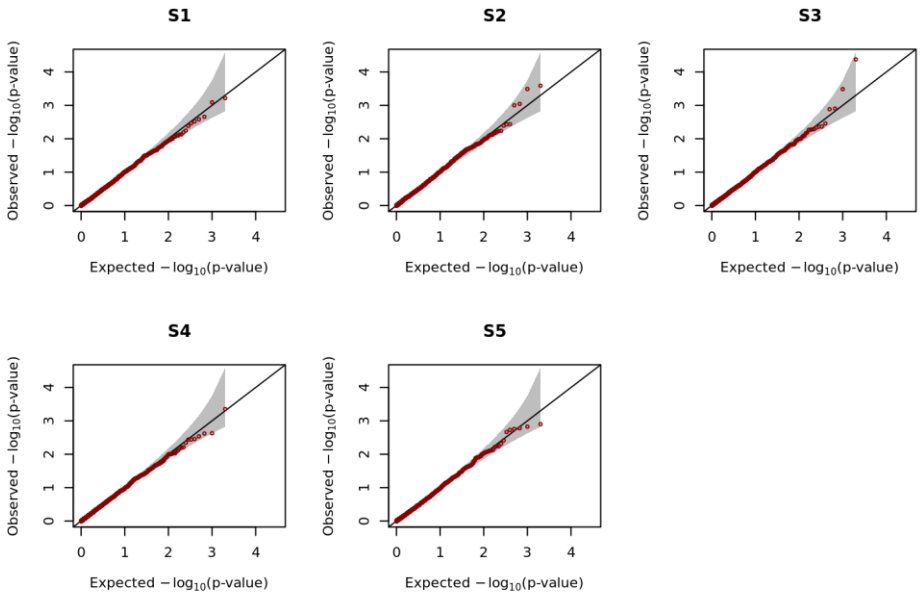


Figure 3.4 Quantile-quantile (QQ) plots of simulated data for scenario 3. I assume that $h^2 = 0.2$ and $q = 0.1$, and scenario 3 was assumed for relatives' family structure. QQ plots were generated from 2,000 replicates

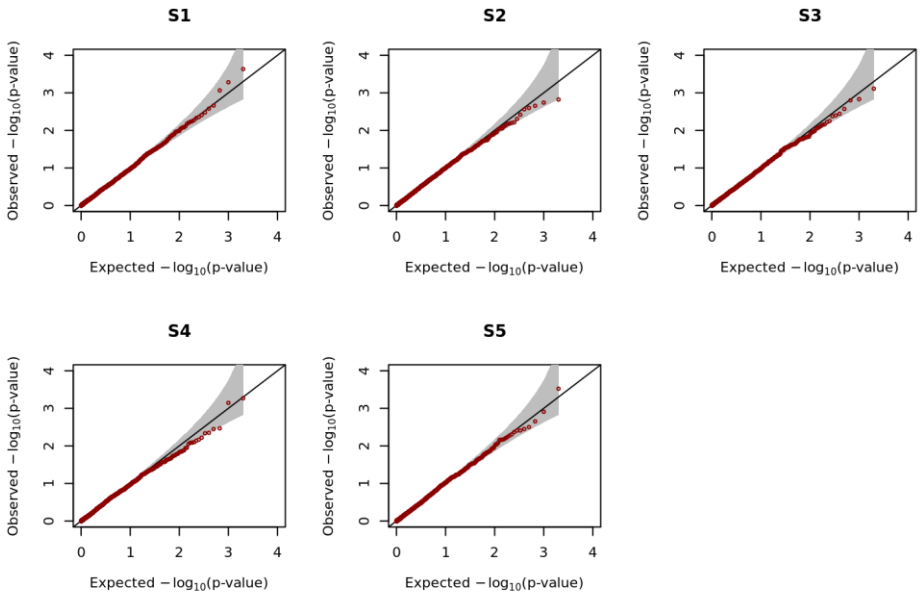


Table 3.2 Empirical type-I error estimates for scenario 2. Scenario 2 was considered for family structures of subjects' relatives. The empirical type-I errors were estimated with 2,000 replicates, and heritabilities were set to be 0.2 and 0.4.

h^2	q	Significance levels	S1[a]	S2[b]	S3[c]	S4[d]	S5[e]	S6[f]
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

[a] S1 : cases and controls were randomly selected from affected and unaffected subjects, respectively

[b] S2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected

[c] S3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)

[d] S4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls

[e] S5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)

[f] S6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls)

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

h^2	q	Significance levels	S1[a]	S2[b]	S3[c]	S4[d]	S5[e]	S6[f]
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

[a] S1 : cases and controls were randomly selected from affected and unaffected subjects, respectively

[b] S2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected

[c] S3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)

[d] S4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls

[e] S5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)

[f] S6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls)

Table 3.4 Empirical power estimates for scenario 1. The empirical power levels were estimated with 2,000 replicates at different levels of significance. I assumed 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[a]	S2[b]	**S3**[c]	S4[d]	S5[e]	**S6**[f]
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**

[a] S1 : cases and controls were randomly selected from affected and unaffected subjects, respectively

[b] S2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected

[c] S3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)

[d] S4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls

[e] S5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)

[f] S6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls)

Table 3.5 Empirical power estimates for scenario 2. The empirical power levels were estimated with 2,000 replicates at different levels of significance. I assumed 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[a]	S2[b]	**S3**[c]	S4[d]	S5[e]	S6[f]
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

[a] S1 : cases and controls were randomly selected from affected and unaffected subjects, respectively

[b] S2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected

[c] S3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)

[d] S4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls

[e] S5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)

[f] S6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls)

Table 3.6 Empirical power estimates for scenario 3. The empirical power levels were estimated with 2,000 replicates at different levels of significance. I assumed 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[a]	S2[b]	**S3**[c]	S4[d]	S5[e]	S6[f]
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

[a] S1 : cases and controls were randomly selected from affected and unaffected subjects, respectively

[b] S2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected

[c] S3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)

[d] S4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls

[e] S5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)

[f] S6 : affected(unaffected) subjects with the largest(smallest) number of affected first-degree relatives were selected as cases(controls)

3.3.3 Robustness of CE to choices of prevalence and heritability

The proposed selection strategy requires heritability and prevalence estimates, and the efficiency of the selection strategy can depend on the accuracy of these estimates. Therefore, I evaluated the sensitivity of the proposed method to misspecification of h^2 and q values using simulated data. I 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 3.7 shows the relative ratio of power estimates for misspecified h^2 and q compared to the results when h^2 and q are correctly specified, with a value of 100 indicating that the power estimates are not affected. Results showed that the effect of misspecification of h^2 and q seems to be almost negligible, at least for the considered simulation models.

Furthermore, ascertained cases and controls remain unchanged as long as the ranks of calculated CEs among cases (and controls) stay the same. I calculated the correlations between orders of true CEs and those with misspecified h^2 and q. Figure 3.5 gives the contour plot of

these correlations. It shows that correlations were always greater than 0.998, even when there were substantial differences between the true and misspecified h^2 and q. Therefore, I can conclude that the rank of CEs remains largely the same, regardless of the values of h^2 and q used.

Table 3.7 Empirical relative power estimates for misspecified heritabilities and prevalences for scenario 3. The empirical power levels were estimated with 2,000 replicates at different levels of significance and the ratios of the power estimates from misspecified (h^2, q) to those from the correctly defined (h^2, q) were calculated as percentage. I 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.

h^2	q	Significance levels	S1[a]	S2[b]	S3[c]	S4[d]	S5[e]
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

[a] S1 : cases and controls were randomly selected from affected and unaffected subjects, respectively

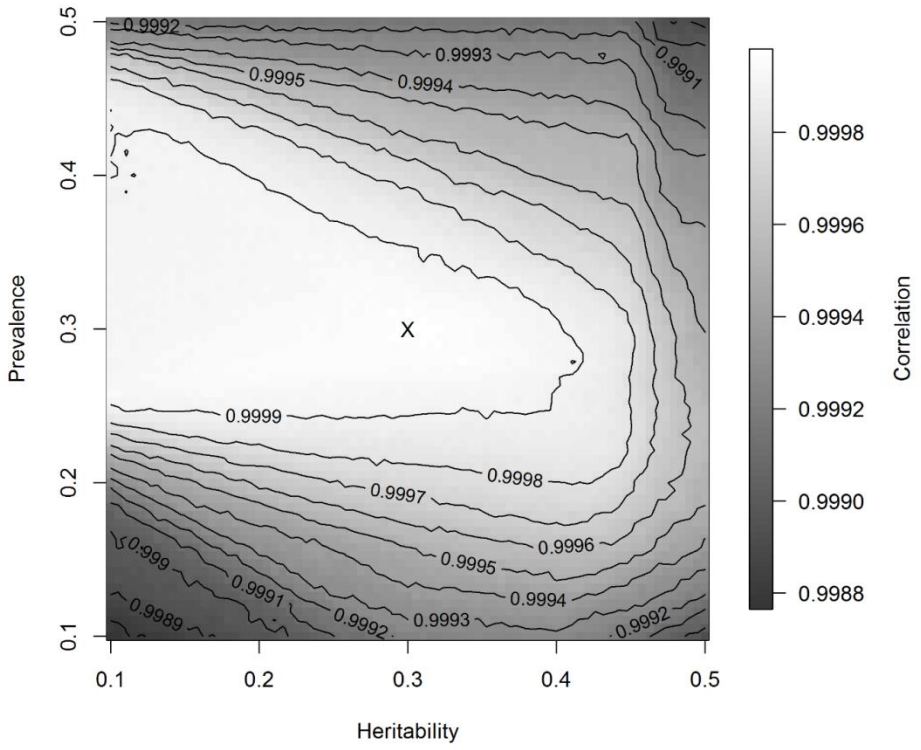
[b] S2 : affected subjects with the highest CEs were selected as cases, and controls were randomly selected

[c] S3 : affected(unaffected) subjects with the highest(lowest) CEs were selected as cases(controls)

[d] S4 : cases were randomly selected, and unaffected subjects with the lowest CEs were selected as controls

[e] S5 : affected(unaffected) subjects with the lowest(highest) CEs were selected as cases(controls)

Figure 3.5 Contour plot for the correlation between orders of conditional expectations (CEs) calculated from true and misspecified (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 were generated from $(h^2, q) = (0.3,0.3)$ and '×' is a point where correlation is exactly 1.



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

3.4.1 The KARE cohort

The KARE cohort was collected to construct an indicator of disease with genetic influences in an attempt to predict the occurrence of various diseases. There are 8,842 participants consisting of 4,183 males and 4,659 females, and they were recruited from two Korean community cohorts, Ansung and Ansan, both in the Gyeonggi Province of South Korea. Participants are 40 to 69 years old. In total, 1,179 subjects were diagnosed as having T2D by a standard guideline (glucose at baseline $\geq$ 126 mg/dL, glucose 120 minutes after the insulin challenge $\geq$ 200 mg/dL, or HbA1c $\geq$ 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, I 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. I also eliminated subjects

with gender inconsistencies, whose identity by state (IBS) was more than 0.8, or whose call rates were less than 95%. As a result, 310,515 SNPs for 8,842 subjects were utilized for GWAS.

3.4.2 The SNUH data

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

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

3.4.3 Association analyses using the pooled data

We used the proposed method to select cases and controls from KARE and SNUH samples for genetic association analyses of T2D. There were a total of 9,523 subjects (8,842 subjects from KARE and 681 subjects from SNUH). I 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 [97], and the heritability of T2D has been reported to be approximately 26% [98]. Therefore, I 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, I selected 1,000 cases and 4,000 controls with S1 and S3. To adjust for population substructure, I calculated a genetic relationship matrix and applied the EIGENSTRAT approach [99]. I obtained the top ten principal component (PC) scores with the largest eigenvalues, and they were included as covariates. I also included sex, age, and squared age as covariates.

3.4.4 Results

We performed genome-wide association study for T2D using the pooled data to compare the performance between selection strategies which I considered in simulation study. The QQ-plots in Figure 3.6 show that GWAS using all subjects and using only the cases and controls ascertained with S1 and S3 preserve the nominal significance levels. Several studies showed that estimates from association analyses with cases and controls selected with family histories of diseases can be inflated [77, 78, 80, 100], and I conducted the other GWAS with permuted phenotypes. Figure 3.7 shows QQ-plots from GWAS with permuted phenotypes and I can conclude that statistical testing is robust against such problems. Figure 3.8 shows Manhattan plots for the analyses, with the genome-wide significance level adjusted by Bonferroni correction (P-value=1.872×10^{-7}) indicated by dashed horizontal lines. The Manhattan plots reveal that the most significant results were obtained from GWAS using all subjects, followed by GWAS using cases and controls ascertained with S3. Table 3.8 shows results for SNPs that were significant in at least one of the GWAS analyses, and it has been reported in some researches that rs10946398, rs7754840, rs9465871, rs7747752, rs9348440, and rs10811661 are associated with T2D. Results showed that GWAS using

cases and controls ascertained with S3 produced more significant SNPs than GWAS using cases and controls ascertained with S1. With the exception of rs10811661, p-values of all SNPs from the S3 GWAS were smaller than those from the S1 GWAS, and the genome-wide significance of SNPs from the S3 GWAS was much larger (Figure 3.9). Therefore, I can conclude that cases and controls ascertained with S3 leads to substantial improvement of power for GWAS.

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

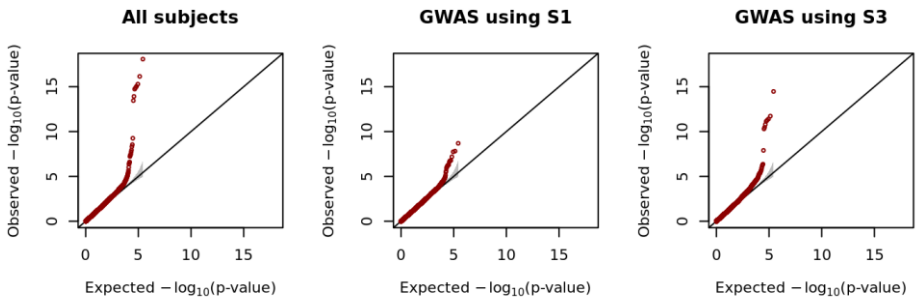


Figure 3.7 Quantile-quantile (QQ) plots for the GWAS with permuted phenotypes.

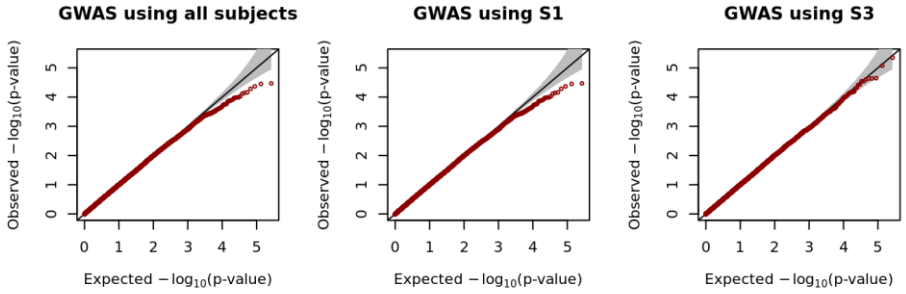


Figure 3.8 Manhattan plots for the results from GWAS of type 2 diabetes.

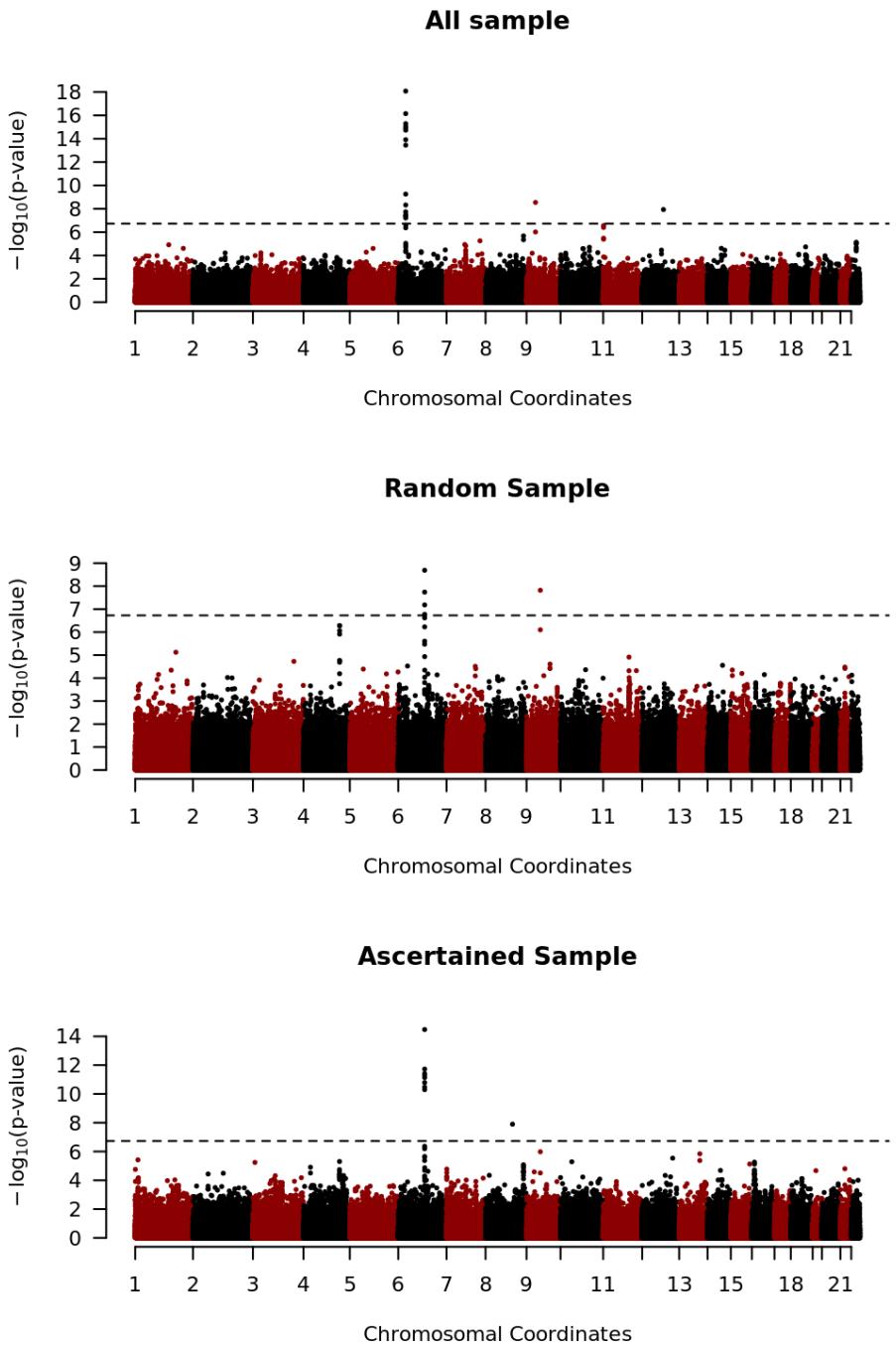
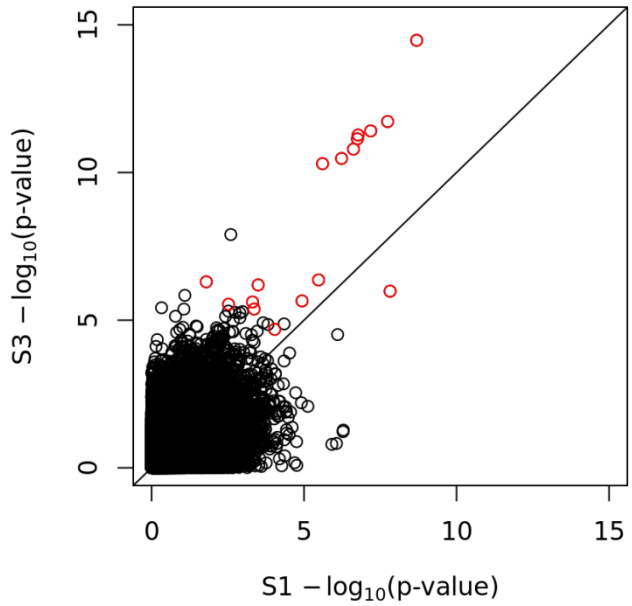


Table 3.8 Results from GWAS. The significance level adjusted by Bonferroni correction is 1.872×10^{-7} and significant SNPs are indicated in bold type.

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

Figure 3.9 Scatter plot for P-values of GWAS of type 2 diabetes using S1 and S3. Red dots indicate significance SNPs when all subjects are used for GWAS.



3.5 Discussion

Many studies have reported that family history of a disease is related to statistical power [77, 78, 80, 100]. 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 [101]. 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, I 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 I 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 [102-106], 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 [89-93], but it is not straightforward to define extreme phenotypes for dichotomous phenotypes. However, subjects with many affected relatives are expected to have higher liability scores, and thus, the presence of a higher number of affected relatives can be used to define extreme phenotypes. Alternatively, if there are continuous phenotypes correlated with the dichotomous phenotypes of interest, they can be utilized to define the extreme phenotypes. Extreme phenotypes can be defined in relation to those continuous phenotypes, and they can be utilized to select subjects. For instance, fasting glucose levels can be used to define extreme phenotypes for type-2 diabetes. Moreover, the use of subjects with extreme phenotypes in GWAS is not the case for selection bias because the choice of subjects is based on phenotype, not on genotype. These approaches can be used with existing software such as MTG2 [107].

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 [108]. The generalized linear model can be understood as a latent variable model if its link function is an inverse function of some cumulative distribution [85]. 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 [85], and the distribution of unknown liability scores can be chosen by comparing several candidate link functions based on the Akaike information criteria [109]. 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 [110]. Alternatively, I can consider the use of generalized estimating equations [111]. However, family histories of diseases have a highly unbalanced structure, which often leads to slow or non-convergence of maximum likelihood estimations or to inflated statistical inferences [112]. 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 [113]. Fortunately, homogeneity tests between cases and controls are statistically valid as long as the estimates of odds ratio are carefully interpreted [113].

Since the introduction of high throughput sequencing technology, substantial reductions in the cost for large-scale genetic association analyses have occurred, and many analyses have been launched to identify loci that show susceptibility. However, large-scale genetic analyses suffer from serious multiple-testing problems, and sequencing remains more expensive than phenotyping. Therefore,

various statistical methods have been investigated to improve the power of testing. Our results reveal that additional statistical power can be achieved in association analyses with careful selection of cases and controls, and that the family history of disease is very useful for this purpose. Furthermore, the family history of disease is often obtained at relatively low costs, and therefore, the proposed method may be a useful strategy for improving the success of genome-wide association analyses.

3.6 Appendix

3.6.1 Calculation of the conditional expectation (CE)

Conditional expectation (CE) is derived with the moment-based approach with minor modifications [86]. If I let $I_A(\cdot)$ be an indicator function and define that

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

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

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

We use the moment-generating function (mgf) of the truncated multivariate normal distribution to calculate the conditional distribution. By definition, I can define the joint probability density function (pdf) of $\mathbf{L}_i$ by

$$f(\mathbf{L}_i) = |2\pi\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{L}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{L}_i\right)$$

where $\boldsymbol{\Sigma}_\mathrm{i} = h^2\boldsymbol{\Psi}_i + (1-h^2)\mathbf{I}_{n_i}$. The conditional pdf of $\mathbf{L}_i$ given $\mathbf{I}_{\mathbf{A}_i}(\mathbf{L}_i) = \mathbf{1}_{n_i+1}$ becomes

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

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

$$m(\mathbf{t}_i) = E\left(e^{\mathbf{t}_i^t \mathbf{L}_i} | \mathbf{I}_{\mathbf{A}_i}(\mathbf{L}_i) = \mathbf{1}_{n_i+1}\right)$$

$$= \frac{1}{\alpha_i |2\pi\boldsymbol{\Sigma}_i|^{1/2}} \int_{\mathbf{A}_i} \exp\left\{-\frac{1}{2}(\mathbf{L}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{L}_i - 2\mathbf{t}_i^t \mathbf{L}_i)\right\} d\mathbf{L}_i$$

where $\mathbf{t}_i = \left(t_{i1}^r, \ldots, t_{in_i}^r, t_i\right)^t$. I let $\boldsymbol{\xi}_i = \boldsymbol{\Sigma}_i \mathbf{t}_i$, and then the exponential term of mgf can be simplified to

$$\exp\left(\frac{1}{2}\mathbf{t}_i^t \boldsymbol{\Sigma}_i \mathbf{t}_i\right) \exp\left\{-\frac{1}{2}(\mathbf{L}_i - \boldsymbol{\xi}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{L}_i - \boldsymbol{\xi}_i)\right\},$$

and mgf becomes

$$m(\mathbf{t}_i) = \frac{\exp(\mathbf{t}_i^t \boldsymbol{\Sigma}_i \mathbf{t}_i / 2)}{\alpha_i |2\pi\boldsymbol{\Sigma}_i|^{1/2}} \int_{\mathbf{A}_i} \exp\left\{-\frac{1}{2}\mathbf{L}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{L}_i\right\} d\mathbf{L}_i.$$

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

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

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

$$F_{ik}^*$$

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

then the CE for subject i can be calculated by

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

3.6.2 Derivation of $F_{ij}(x)$

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

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

If I 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, I know that $\mathbf{L}_i^r | L_i = x$ is normally distributed with

$$E(\mathbf{L}_i^r | L_i = x) = \mathbf{\Sigma}_i^{rI} x \text{ and } \text{var}(\mathbf{L}_i^r | L_i = x) = \mathbf{\Sigma}_i^r - \mathbf{\Sigma}_i^{rI} (\mathbf{\Sigma}_i^{rI})^t.$$

Therefore, the multivariate marginal pdf of L_i becomes

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

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* [114].

Chapter 4

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

4.1 Introduction

Phenotypes are affected both by environmental factors and genes, and family members are expected to possess similar phenotypes due to their genetic similarity. Heritability was defined to quantify phenotypic similarity attributable to heritable components, and this concept has been widely used to understand the genetic architecture of phenotypes [115]. For example, heritability can be used to compare the

importance of genetic components among different phenotypes. Additionally, if large-scale genetic data are available, genetic correlation matrices can be estimated [116]. These data can then be incorporated into a linear mixed model to provide SNP heritability estimation. SNP heritability provides information regarding the relative proportion of variance attributable to the genotyped SNPs, and this technique can be used to identify the degree of missing heritability.

Estimation of broad-sense heritability requires the study of bilinear relatives such as sibling or monozygotic twins, and in practice, narrow-sense heritability has often been utilized. Narrow-sense heritability is defined as the proportion of the total phenotypic variation explained by additive genetic effects [115]. Various methods have been developed for estimating the heritability of continuous traits. For example, restricted maximum likelihood methods based on the linear mixed model (LMM) [22, 117, 118] or polygenic score methods [119] can be used for estimating the heritability of continuous traits. For dichotomous traits, generalized linear mixed models or Liability Threshold Models have been often utilized [21, 120]. The Liability Threshold Model assumes there are unobserved continuous liability scores, and subjects are affected if they exceed a certain threshold [16, 22, 121, 122].

In this study, I focus on heritability estimation of dichotomous phenotypes. There are multiple factors which can bias variance estimation of dichotomous traits. In particular, family-based samples are typically analyzed using probands. The term proband refers to instances when family members are brought into a study as a result of other family members already enrolled in the study. Multiple reports indicate that proband analysis can produce substantial bias in variance estimates [22, 123, 124]. For example, if phenotypes are rare and families are randomly selected, the number of affected individuals is often very small. Therefore families are ascertained through the use of affected probands. In such instances, the majority of the relatives may be unaffected unless the size of the family is very large, and negative correlation can be observed because probands are affected while their relatives are unaffected. Several approaches have been proposed to adjust for such bias. GCTA adjusts estimated heritabilities by assuming that the level of ascertainment bias is same among individuals [22]; however, families are ascertained with probands and the effect of ascertainment bias is heterogeneous according to familial relationship [124]. For example, ascertainment bias for grandparents of the proband is expected to be approximately half that of the parents.

Here, I developed a new method to estimate heritability based on the Liability Threshold Model for binary traits (LTMH) which can be applied to the extended pedigree structure. Using the Expectation-Maximization (EM) algorithm, the proposed method jointly estimates maximum likelihood estimators (MLE) for heritability and coefficients of covariates [14]. Furthermore, the proposed method maximizes the conditional likelihood of disease statuses of probands via a conditional EM (CEM) algorithm [125], and ascertainment bias can be adjusted. I also developed a conditional expected score test (CEST) to determine if heritability is equal to zero. Extensive simulation studies demonstrated that heritability estimates obtained from the proposed methods are generally unbiased even for the ascertained family-based samples. Estimates from GCTA are unbiased for randomly selected families, but the bias turns out to be substantial for ascertained families. Also I found that the CEST for heritability was statistically conservative, but it could achieve reasonable statistical power estimates. Finally, I used the proposed method to estimate the heritability of type-2 diabetes (T2D) using ascertained family-based samples from Korean families, and those estimates confirmed the practical value of our proposed methods.

4.2 Materials and Methods

4.2 1 Notations and Disease Model

We assume that there are n independent families and family i has n_i family members ($i = 1, \ldots, n$). I consider the Liability Threshold Model, and assume dichotomous phenotypes are determined by the unobserved continuous liability score. The liability score of subject j in family i is denoted by L_{ij}, and they are determined by summing the environmental/genetic effects, polygenic effects, and random error. The covariates including environmental/genetic effects for subject j in family i are denoted by $\mathbf{X}_{ij}$, and I assumed that covariates are standardized. In this article, I assumed there are p covariates. The random effects, including polygenic effect and random error for subject j in family i, are denoted by U_{ij}. The vector forms of those components for family i are denoted by:

$$\mathbf{L_i} = \begin{pmatrix} L_{i1} \\ \vdots \\ L_{in_i} \end{pmatrix}, \mathbf{X_i} = \begin{pmatrix} \mathbf{X}_{i1} \\ \vdots \\ \mathbf{X}_{in_i} \end{pmatrix} \text{ and } \mathbf{U_i} = \begin{pmatrix} U_{i1} \\ \vdots \\ U_{in_i} \end{pmatrix}.$$

Liability scores of family members are usually correlated, and I assumed that those are normally distributed as follows:

$$\mathbf{L}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i, \mathbf{L}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$$

where $\mathbf{U}_i \sim MVN(\mathbf{0}, \mathbf{\Sigma}_i)$. I denote $\mathbf{\Phi}_i$ to be the kinship coefficient matrix multiplied by two, and $\mathbf{I}_w$ is the $w \times w$ dimensional identity matrix. Under the polygenic model using additivity of genetic effects across loci and linkage equilibrium among loci, I can get:

$$\mathbf{\Sigma}_i = \sigma_a^2 \mathbf{\Phi}_i + \sigma_d^2 \mathbf{V}_{di} + \sigma_h^2 \mathbf{V}_{hi} + \sigma_{a,d} \mathbf{V}_{adi} + \sigma_e^2 \mathbf{I}_{n_i}$$

where σ_a^2, σ_d^2 and σ_e^2 are the variances of additive, dominant, and environmental effects in the population, and σ_h^2 and $\sigma_{a,d}$ are the dominant genetic variance and the covariance of additive and dominant effects in the homozygous population, respectively [126-128]. $\mathbf{V}_{di}$, $\mathbf{V}_{hi}$ and $\mathbf{V}_{adi}$ are the functions of the condensed coefficients of identity [128]. For simplicity, I assume that all variance components other than σ_a^2 and σ_e^2 are zero, and the sum of σ_a^2 and σ_e^2 is equal to one. If I denote heritability as $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, then the variance-covariance matrix of $\mathbf{\Sigma}_i$ is expressed by

$$\mathbf{\Sigma}_i = h^2 \mathbf{\Phi}_i + (1 - h^2) \mathbf{I}_{n_i}.$$

The dichotomous phenotypes for subject j in family i are denoted by Y_{ij} and these values are coded as 1 for cases and 0 for controls. Phenotype vector for family i is denoted by:

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}.$$

In a Liability Threshold Model, Y_{ij} is determined by L_{ij}, and if L_{ij} is larger than a certain threshold value c, Y_{ij} becomes 1, and otherwise it becomes 0. c can be determined from the prevalence of the diseases as c should be the inverse of the cumulative distribution function of the prevalence. For each observed Y_{ij}, I can infer the range of the corresponding L_{ij}, (a_{ij}, b_{ij}). For example, if $Y_{ij} = 0$, then L_{ij} is bounded by $(-\infty, c)$, and otherwise, L_{ij} is bounded by (c, ∞). The lower and upper bounds of the liability for the family i are denoted by:

$$\mathbf{a_i} = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{in_i} \end{pmatrix} \text{ and } \mathbf{b_i} = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{in_i} \end{pmatrix}.$$

Based on above notations, all subjects can be expressed in the following vector forms:

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 \\ \vdots \\ \mathbf{L}_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix}, \mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_n \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{pmatrix},$$

$$\mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_\mathrm{n} \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{pmatrix}.$$

Under those notations, I assumed that $\mathbf{L}$ follows multivariate normal distribution with mean $\mathbf{X\beta}$ and variance-covariance matrix $\mathbf{\Sigma}$ which exist in a block diagonal matrix consisting of $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_n$.

4.2.2 Heritability Estimation using the EM Algorithm

The EM (Expectation-Maximization) algorithm [14] was used to estimate h^2 based on the complete data consisting of observed phenotypes, **Y**, and unobserved liabilities, **L**. The joint probability density function (pdf) of the complete data can be decomposed into the marginal pdf of **L** and the conditional pdf of **Y** given that **L** has the support of (**a, b**). This can be formulated as:

$$f(\mathbf{Y}, \mathbf{L}) = f(\mathbf{Y}|\mathbf{L})f(\mathbf{L}) = f(\mathbf{L})I(\mathbf{a} < \mathbf{L} < \mathbf{b}).$$

If I define the parameters of interest as $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, h^2)^t$, then the log-likelihood of the complete data will be the sum of the log-likelihoods for each family as follows:

$$l(\boldsymbol{\theta};\ \mathbf{Y}, \mathbf{L}) = \sum_{i=1}^{n} \Big[-\frac{n_i}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{L}_i - \mathbf{X}_i\boldsymbol{\beta})^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{L}_i - \mathbf{X}_i\boldsymbol{\beta}) \Big].$$

In the E-step of the EM algorithm, the conditional expectation of **L** given **Y** was taken to the $l(\boldsymbol{\theta};\ \mathbf{Y}, \mathbf{L})$, where the estimates for the parameters of the previous iteration were used. If I assume that the *k*th iteration has been performed and denote the estimates for the parameters at the *k*th iteration as $\boldsymbol{\theta}^{(k)}$, then the conditional expectation $Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)$ will be

$$Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right) = E_{\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}}[l(\boldsymbol{\theta};\, \mathbf{Y}, \mathbf{L})] = \sum_{i=1}^{n} E_{\mathbf{L}_i|\mathbf{Y}_i,\boldsymbol{\theta}^{(k)}}[l_i(\boldsymbol{\theta};\, \mathbf{Y}_i, \mathbf{L}_i)]$$

$$= \sum_{i=1}^{n} Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)$$

and

$$Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right) = -\frac{n_i}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i|$$

$$-\frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i^{(k)}\right) - 2\boldsymbol{\beta}^t\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i^{(k)} + \boldsymbol{\beta}^t\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\boldsymbol{\beta}\right]$$

where $\mathbf{A}_i^{(k)} = E_{\mathbf{L}_i|\mathbf{Y}_i,\boldsymbol{\theta}^{(k)}}(\mathbf{L}_i\mathbf{L}_i^t)$ and $\mathbf{B}_i^{(k)} = E_{\mathbf{L}_i|\mathbf{Y}_i,\boldsymbol{\theta}^{(k)}}(\mathbf{L}_i)$. $\mathbf{A}_i^{(k)}$ and $\mathbf{B}_i^{(k)}$ are equal to the first moment and the second moment of the multivariate truncated normal, respectively. R package *tmvtnorm* was utilized for calculation [86].

In the M-step of the EM algorithm, I maximize $Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)$ with respect to $\boldsymbol{\theta}$. Since $Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)$ is the concave function, I can find the maximizer by solving for $\partial Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)/\partial\boldsymbol{\theta} = 0$. The partial derivative with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i^{(k)} - \sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\boldsymbol{\beta} \qquad (1)$$

and, $\boldsymbol{\beta}^{(k)}(h^2)$ which satisfies $\partial Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)/\partial\boldsymbol{\beta} = 0$ becomes

$$\boldsymbol{\beta}^{(k)}(h^2) = \left(\sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i^{(k)}\right).$$

To emphasize that the root is the function of h^2, it was denoted by $\boldsymbol{\beta}^{(k)}(h^2)$. Unfortunately, there is no closed form of the root in which $\partial Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)/\partial h^2 = 0$, and generalized EM algorithms were applied. $\boldsymbol{\theta}^{(k)}$ was updated using a Newton-Raphson algorithm [129]. After I obtained the maximizer of $Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)$ during the maximization step, I updated $\boldsymbol{\theta}^{(k)}$ to $\boldsymbol{\theta}^{(k+1)}$ and repeated the EM steps until convergence. The detailed algorithm is provided in Appendix (A).

Note that $\widehat{\boldsymbol{\beta}}$ is the unbiased estimator of $\boldsymbol{\beta}$ and it can be easily proven by

$$E_{\mathbf{Y}_i}\left(\mathbf{B}_i^{(m)}\right) = E_{\mathbf{Y}_i}\left(E_{\mathbf{L}_i|\mathbf{Y}_i,\boldsymbol{\theta}^{(m)}}(\mathbf{L}_i)\right) = E_{\mathbf{L}_i}(\mathbf{L}_i) = \mathbf{X}_i\boldsymbol{\beta}$$

assuming I obtained $\widehat{\boldsymbol{\beta}}$ after *m* iterations [100].

4.2.3 Lagrangian Multiplier and Karush-Kuhn-Tucker Condition

Unlike $\boldsymbol{\beta}$, the parameter space of h^2 is restricted to $\Theta_{h^2} = \{h^2 : 0 \leq h^2 \leq 1\}$, and the objective function should be maximized under the restriction as follows:

$$\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right) \text{ subject to } 0 \leq h^2 \leq 1.$$

This objective function can be maximized using the method of Lagrange multiplier [130] under Karush-Kuhn-Trucker (KKT) conditions [131]. The constraint is equivalent to $-h^2 \leq 0$ and $h^2 - 1 \leq 0$, and by the Lagrangian multiplier, the object function becomes

$$Q^*\left(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\theta}^{(k)}\right) = Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right) + \lambda_1 h^2 - \lambda_2(h^2 - 1)$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^t$. I can find the solution that maximizes $Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)$ subject to $0 \leq h^2 \leq 1$ by finding $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ satisfying the following three conditions known as KKT conditions:

1) *Stationarity* : $\partial Q^*\left(\boldsymbol{\theta}, \boldsymbol{\lambda}\middle|\boldsymbol{\theta}^{(k)}\right)/\partial\boldsymbol{\theta} = \mathbf{0}$,

2) *Complementary slackness* : $\lambda_1 h^2 = 0$ and $\lambda_2(1 - h^2) = 0$,

3) *Dual feasibility* : $\lambda_i \geq 0$ for $i = 1,2$.

More specifically, for the *Stationarity* condition, $\partial Q^*\left(\boldsymbol{\theta},\boldsymbol{\lambda}\middle|\boldsymbol{\theta}^{(k)}\right)/\partial\boldsymbol{\beta}$ is identical to $\partial Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)/\partial\boldsymbol{\beta}$, providing that $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(k)}(h^2)$. Replacing $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^{(k)}(h^2)$, I get

$$\left.\frac{\partial Q^*\left(\boldsymbol{\theta},\boldsymbol{\lambda}\middle|\boldsymbol{\theta}^{(k)}\right)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}(h^{2*}),h^2=h^{2*}}$$

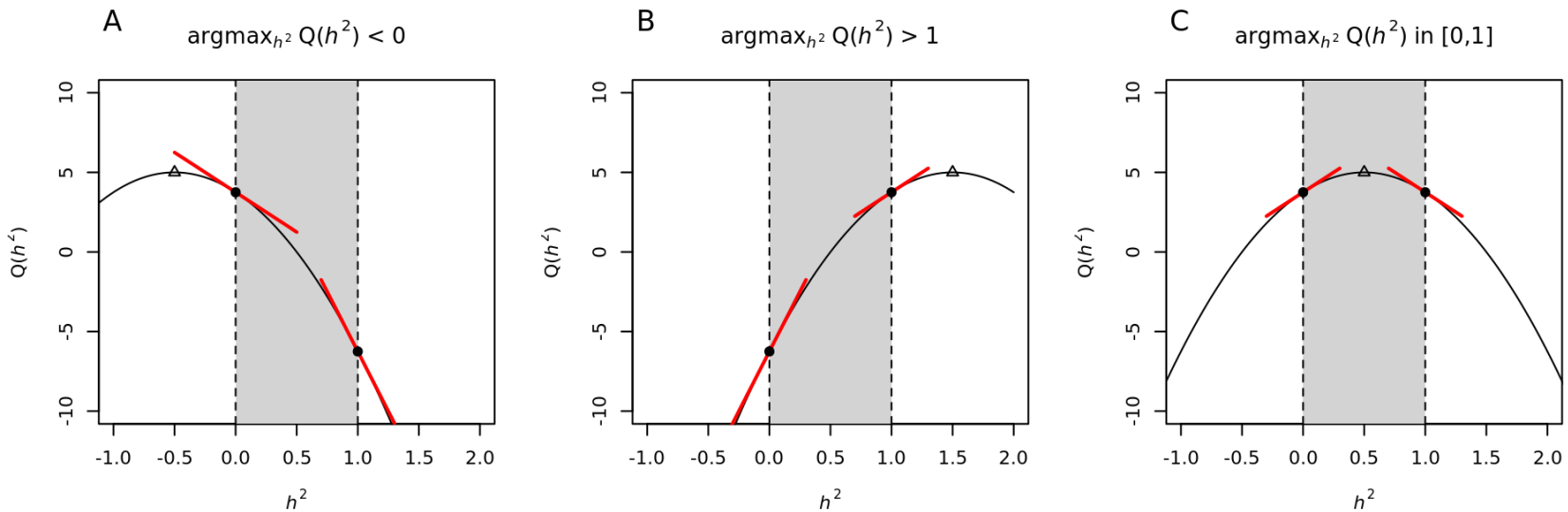
$$= \left.\frac{\partial Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}(h^{2*}),h^2=h^{2*}} + \lambda_1 - \lambda_2 = 0,$$

and it is equivalent to

$$\left.\frac{\partial Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}(h^{2*}),h^2=h^{2*}} = -\lambda_1 + \lambda_2.$$

Note that to the left of this equation is a function of h^{2*}, denoted by $g^{(k)}(h^{2*})$. Applying *Complementary slackness* conditions to the above equation, $(\lambda_1,\lambda_2,h^2)$ becomes $(0,0,h^2)$, $(\lambda_1,0,0)$, or $(0,\lambda_2,1)$. If I assume $h^2 = 0$ and $\lambda_2 = 0$, then $g^{(k)}(0) = -\lambda_1$ and it will be non-positive if the assumptions are met by the *Dual feasibility* condition. Similarly, when $h^2 = 1$ and $\lambda_1 = 0$ are assumed, $g^{(k)}(1) = \lambda_2$ and it will be non-negative if the assumptions are satisfied. If none of these assumptions are met, λ_1 and λ_2 are automatically zero, and thus optimization can be done without any restrictions on h^2. This concept is illustrated in Figure 4.1.

Figure 4.1 Illustration of KKT condition using a toy example. The exemplary concave function $Q(h^2)$ was created to enable determination of the optimal value that maximizes $Q(h^2)$ within the parameter space. The parameter h^2 can be between zero and one, and the parameter space for this value is grayed out. (A) If the value that maximizes $Q(h^2)$ is negative, the tangent slopes at both zero and one will be negative. A tangent slope that is negative at one violates the KKT conditions, however, a negative tangent slop at zero satisfies the KKT conditions, so the maximizer within the parameter space is zero. (B) When the value which maximizes $Q(h^2)$ is greater than 1, the optimal value is one since positive tangent slope at one meets the KKT conditions. (C) When the maximizer is located in the parameter space, tangent slopes at both boundaries of the parameter space do not satisfy the KKT conditions. Therefore, restrictions do not affect the result of optimization.


A
argmax$_{h^2}$ Q(h^2) < 0
B
argmax$_{h^2}$ Q(h^2) > 1
C
argmax$_{h^2}$ Q(h^2) in [0,1]
Q(h^2)
h^2
10
5
0
-5
-10
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

4.2.4 Ascertainment Bias-corrected Heritability Estimation

Ascertainment of each family is conducted using probands, and statistical inferences about heritability may be misleading unless ascertainment is correctly adjusted. I assume the first family member in each family is a proband, and the other $n_i - 1$ family members are non-probands. To distinguish probands and non-probands, I added superscripts *P* and *NP*, respectively. Vectors for liabilities, covariates, phenotypes, and bounds of liabilities for non-probands in family *i* are denoted by:

$$\mathbf{L}_i^{NP} = \begin{pmatrix} L_{i2}^{NP} \\ \vdots \\ L_{in_i}^{NP} \end{pmatrix}, \mathbf{X}_i^{NP} = \begin{pmatrix} \mathbf{X}_{i2}^{NP} \\ \vdots \\ \mathbf{X}_{in_i}^{NP} \end{pmatrix}, \mathbf{Y}_i^{NP} = \begin{pmatrix} Y_{i2}^{NP} \\ \vdots \\ Y_{in_i}^{NP} \end{pmatrix}, \mathbf{a}_i^{NP} = \begin{pmatrix} a_{i2}^{NP} \\ \vdots \\ a_{in_i}^{NP} \end{pmatrix} \quad \text{and}$$

$$\mathbf{b}_i^{NP} = \begin{pmatrix} b_{i2}^{NP} \\ \vdots \\ b_{in_i}^{NP} \end{pmatrix}.$$

Similarly, those variables pertaining to a proband in family *i* are defined as L_i^P, $\mathbf{X}_i^P$, Y_i^P, a_i^P and b_i^P, respectively. Liability vectors for probands and non-probands across entire families are denoted by:

$$\mathbf{L}^P = \begin{pmatrix} L_1^P \\ \vdots \\ L_n^P \end{pmatrix}, \mathbf{L}^{NP} = \begin{pmatrix} \mathbf{L}_1^{NP} \\ \vdots \\ \mathbf{L}_n^{NP} \end{pmatrix} \text{ and } \mathbf{L} = \begin{pmatrix} \mathbf{L}^P \\ \\ \mathbf{L}^{NP} \end{pmatrix},$$

and vectors for other variables are also similarly defined.

To adjust for the effects of ascertainment on heritability estimates, I estimated parameters using the following conditional likelihood:

$$f(\mathbf{Y}^{NP}|\mathbf{Y}^{P};\boldsymbol{\theta}) = \frac{f(\mathbf{Y};\boldsymbol{\theta})}{f(\mathbf{Y}^{P};\boldsymbol{\theta})}.$$

If I assume $l(\boldsymbol{\theta};\ \mathbf{Y}) = \log f(\mathbf{Y};\boldsymbol{\theta})$, the log of the conditional likelihood is $l(\boldsymbol{\theta};\ \mathbf{Y}) - l(\boldsymbol{\theta};\ \mathbf{Y}^{P})$. The objective function of the EM algorithm is a global lower bound for the log-likelihood [132], and if I assume the lower bound $\mathcal{F}(\boldsymbol{\theta})$ for $l(\boldsymbol{\theta};\ \mathbf{Y})$ and the upper bound $\mathcal{G}(\boldsymbol{\theta})$ for $l(\boldsymbol{\theta};\ \mathbf{Y}^{P})$, then the global lower bound can be obtained by:

$$\log f(\mathbf{Y}^{NP}|\mathbf{Y}^{P};\boldsymbol{\theta}) \geq\ \mathcal{F}(\boldsymbol{\theta}) - \mathcal{G}(\boldsymbol{\theta}).$$

At $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, $\mathcal{F}(\boldsymbol{\theta})$ can be obtained by:

$$\mathcal{F}(\boldsymbol{\theta}) = E_{\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}}\big(l(\boldsymbol{\theta};\ \mathbf{Y},\mathbf{L})\big) + H\left(f\big(\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}\big)\right),$$

where $H(\cdot)$ is the entropy. The upper bound $\mathcal{G}(\boldsymbol{\theta})$ for $l(\boldsymbol{\theta};\ \mathbf{Y}^{P})$ can be defined as $l(\boldsymbol{\theta};\ \mathbf{Y}^{P}) + \text{constant}$ [125]. Therefore, the global lower bound of the log-likelihood at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ becomes:

$$\mathcal{F}(\boldsymbol{\theta}) - \mathcal{G}(\boldsymbol{\theta}) = E_{\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}}\big(l(\boldsymbol{\theta};\ \mathbf{Y},\mathbf{L})\big) - l(\boldsymbol{\theta};\ \mathbf{Y}^{P}) + \text{constant}.$$

We assume probands are independent of each other, and proband i was randomly selected from the population with the probability μ_i. Then, $l(\boldsymbol{\theta};\ \mathbf{Y}^{P})$ is simply given by:

$$l(\boldsymbol{\beta};\ \mathbf{Y}^P) = \sum_{i=1}^{n} l(\boldsymbol{\beta}; Y_i^P) = \sum_{i=1}^{n} [Y_i^P \alpha_i - \log(1 + e^{\alpha_i})]$$

where $\alpha_i = \log \frac{\mu_i}{1-\mu_i}$. Here μ_i is formulated as a function of the cumulative distribution function of the standard normal, $\Phi(\cdot)$, by:

$$\mu_i = E(Y_i^P) = \Pr(Y_i^P = 1) = \Pr(L_i^P > c) = 1 - \Phi(c - \mathbf{X}_i^P \boldsymbol{\beta}).$$

The MLE values for $\boldsymbol{\theta}$ are obtained by iteratively maximizing the objective function until convergence, and the detailed algorithm for maximization is provided in Appendix (B).

4.2.5 Conditional Expected Score Tests

$\boldsymbol{\beta}$ and h^2 are required to parameterize the relationship between covariates and $\mathbf{Y}$ at the unobserved liability scale, and I consider the conditional expected score test (CEST) [14, 133, 134] because:

$$\frac{\partial l(\boldsymbol{\theta};\, \mathbf{Y})}{\partial \boldsymbol{\theta}} = E_{\mathbf{L}|\mathbf{Y}}\left[\frac{\partial l(\boldsymbol{\theta};\, \mathbf{Y},\mathbf{L})}{\partial \boldsymbol{\theta}}\right].$$

For simplicity, I assumed that the prevalence is correctly specified and samples are randomly selected. The conditional expected score based on the complete data for family i is:

$$\mathbf{S}_i = E_{\mathbf{L}|\mathbf{Y}}\begin{bmatrix}\dfrac{\partial l_i(\boldsymbol{\theta};\, \mathbf{Y},\mathbf{L})}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial l_i(\boldsymbol{\theta};\, \mathbf{Y},\mathbf{L})}{\partial h^2}\end{bmatrix}$$

$$= \begin{bmatrix}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i - \mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\boldsymbol{\beta} \\ -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i}\right)\right) - \frac{1}{2}\mathrm{tr}(\mathbf{C}_i\mathbf{A}_i) + \boldsymbol{\beta}^t\mathbf{X}_i^t\mathbf{C}_i\left(\mathbf{B}_i - \frac{1}{2}\mathbf{X}_i\boldsymbol{\beta}\right)\end{bmatrix}$$

where $\mathbf{A}_i = E_{\mathbf{L}|\mathbf{Y}}(\mathbf{L}_i\mathbf{L}_i^t)$, $\mathbf{B}_i = E_{\mathbf{L}|\mathbf{Y}}(\mathbf{L}_i)$ and $\mathbf{C}_i = \partial\boldsymbol{\Sigma}_i^{-1}/\partial h^2$. Note that $\mathbf{A}_i$ and $\mathbf{B}_i$ are also a function of $\boldsymbol{\theta}$. If I assume $\mathbf{S}_{\boldsymbol{\beta}i}$ and $S_{h^2 i}$ denote $E_{\mathbf{L}|\mathbf{Y}}[\partial l_i(\boldsymbol{\theta};\, \mathbf{Y},\mathbf{L})/\partial\boldsymbol{\beta}]$ and $E_{\mathbf{L}|\mathbf{Y}}[\partial l_i(\boldsymbol{\theta};\, \mathbf{Y},\mathbf{L})/\partial h^2]$, respectively, then the score statistics can be obtained by:

$$\mathbf{S} = \begin{pmatrix}\mathbf{S}_{\boldsymbol{\beta}}^t & S_{h^2}\end{pmatrix}^t \text{ where } \mathbf{S}_{\boldsymbol{\beta}} = \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\beta}i}, \text{ and } S_{h^2} = \sum_{i=1}^n S_{h^2 i}.$$

The variance-covariance matrix of $\mathbf{S}$ is calculated using the observed Fisher information matrix [135, 136]. The observed Fisher information matrix is given by:

$$\hat{I}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\mathbf{S}_i \mathbf{S}_i^t) - \frac{1}{n}\left(\sum_{i=1}^{n} \mathbf{S}_i\right)\left(\sum_{i=1}^{n} \mathbf{S}_i^t\right)$$

and it is equivalent to:

$$\hat{I}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{i}_{\boldsymbol{\beta}} & \boldsymbol{i}_{\boldsymbol{\beta} h^2} \\ \boldsymbol{i}_{h^2 \boldsymbol{\beta}} & i_{h^2} \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{i=1}^{n} \left(\mathbf{S}_{\boldsymbol{\beta} i} \mathbf{S}_{\boldsymbol{\beta} i}^t\right) - \mathbf{S}_{\boldsymbol{\beta}} \mathbf{S}_{\boldsymbol{\beta}}^t / n & \sum_{i=1}^{n} \left(\mathbf{S}_{\boldsymbol{\beta} i} S_{h^2 i}\right) - \mathbf{S}_{\boldsymbol{\beta}} S_{h^2} / n \\ \sum_{i=1}^{n} \left(S_{h^2 i} \mathbf{S}_{\boldsymbol{\beta} i}\right) - S_{h^2} \mathbf{S}_{\boldsymbol{\beta}}^t / n & \sum_{i=1}^{n} \left(S_{h^2 i}^2\right) - S_{h^2}^2 / n \end{pmatrix}.$$

Therefore, if I assume p to be the dimension of $\boldsymbol{\beta}$, and $\widehat{h^2}$ and $\widehat{\boldsymbol{\beta}}$ are MLEs, I can provide the following statistics [135, 136]:

$$\mathbf{S}_{\boldsymbol{\beta}}^{\boldsymbol{t}} \left\{ \boldsymbol{i}_{\boldsymbol{\beta}} - \boldsymbol{i}_{\boldsymbol{\beta} \widehat{h^2}} i_{\widehat{h^2}}^{-1} \boldsymbol{i}_{\widehat{h^2} \boldsymbol{\beta}} \right\}^{-\mathbf{1}} \mathbf{S}_{\boldsymbol{\beta}} \sim \chi^2 (df = p) \text{ under } H_0 \colon \boldsymbol{\beta} = \mathbf{0}.$$

To test if $H_0 \colon h^2 = 0$, the likelihood is maximized at $h^2 = 0$ with 50% probability and at the positive real number at 50% probability under H_0. Thus I consider:

$$\mathbf{S}_{h^2}^{\boldsymbol{t}} \left\{ i_{h^2} - \boldsymbol{i}_{h^2 \widehat{\boldsymbol{\beta}}} \boldsymbol{i}_{\widehat{\boldsymbol{\beta}}}^{-1} \boldsymbol{i}_{\widehat{\boldsymbol{\beta}} h^2} \right\}^{-\mathbf{1}} \mathbf{S}_{h^2} \sim \frac{\mathbf{1}}{\mathbf{2}} \cdot \mathbf{0} + \frac{\mathbf{1}}{\mathbf{2}} \cdot \chi^2 (df = 1)$$

under $H_0 \colon h^2 = 0$.

4.2.6 Simulation studies

Simulation studies were conducted under two different scenarios where families were either randomly selected (scenario 1) or ascertained with probands (scenario 2).

For scenario 1, 500 families were randomly generated. For scenario 2, 50,000 families for each replicate were initially generated. Then, 500 probands were selected from affected individuals, and their family members were determined. For both scenarios, I considered nuclear families and the number of siblings at 1, 2, 3 and 4 with proportions of 0.2, 0.3, 0.3 and 0.2, respectively. Liabilities were determined through summation of major genetic effects, polygenic effects, and random errors. Sums of polygenic effects and random errors were generated using multivariate normal distribution with heritability values of 0.05, 0.2 and 0.4. The main genetic effects were obtained using the product of β and the number of disease alleles. Disease allele frequency was assumed to be 0.2, and genotype frequencies were obtained under HWE. Founder genotypes for each family were generated from B (2, 0.2), and non-founder genotypes were obtained by examining Mendelian transmission. β was obtained by h_a^2 and disease allele frequency (p) using the following equation:

$$h_a^2 = \frac{2\beta^2 p(1-p)}{2\beta^2 p(1-p) + 1}.$$

h_a^2 was assumed to be 0.005 and β was 0.1253. Once liabilities were generated, they were considered affective if they were larger than the threshold c. Otherwise, they were considered non-affective. c was chosen to maintain the assumed prevalences (q). The R code for generating the simulation data can be downloaded from http://healthstat.snu.ac.kr/software/LTMH.

The performance of our experimental method was evaluated using 2,000 replicates exhibiting various combinations of heritabilities (h^2) and prevalences (q). For evaluation of statistical testing of β, the q were set at 0.1 or 0.2, and h^2 was assumed to be 0.2 or 0.4. For evaluation of statistical testing for h^2, I assumed q = 0.05, 0.1 or 0.2 and h^2 = 0, 0.2 and 0.4. All results were compared to GCTA results for each scenario.

4.2.7 Application for Family-based Samples of Type-2 Diabetes

The proposed method was applied to the cross-sectional study of T2D patients conducted by Seoul National University Hospital in Korea. T2D patients were diagnosed according to the World Health Organization criteria for T2D [137]. The study preferentially included T2D patients with a positive family history of T2D in first-degree relatives, and 681 probands were recruited. Family histories of T2D were obtained based on the memory of probands, but the study excluded relatives who were positive for the 75-g oral glucose tolerance test. Subjects of unknown age were also excluded, and 4,149 non-probands, including 1,115 T2D patients and 648 affected probands, remained. For our analyses, the effect of age was adjusted through use as a covariate, and standardized age was incorporated into final analyses. The prevalence of T2D was set at 10.9% [138], and the heritability of T2D was estimated using our experimental method adjusted for ascertainment bias.

4.2.8 Application for GWAS of S-LAM

We applied CEST for GWAS of case-control study of S-LAM disease. S-LAM patients were collected from 2010 to 2014 from 14 countries and DNA samples for 479 S-LAM patients were genotyped with the Infinium OmniExpress-24 v1.2 BeadChip. I excluded 34 non-white S-LAM patients, and finally 445 S-LAM patients were used to GWAS as cases with 716,503 SNPs. For controls of GWAS, I used 1,261 healthy female from the COPDGene Consortium. I filtered out all SNPs whose P-value of HWE test is less than 1×10^{-5}, MAF is less than 0.05 or genotype call rates were less than 95%. I also excluded all subjects whose genotype call rates were less than 95% or identity-by-states were larger than 80% with any other subject. To compare statistical power of CEST to the conditional logistic regression (CLR), I matched each cases with two controls using age of enroll and two PC scores. Each pair of one case and two controls was regarded as if a family having relatedness structure of genetic relationship matrix. Finally, 426 S-LAM cases and 852 cases were included for GWAS with 549,599 SNPs. Detailed QC procedure is described in Chapter 2.2.2 (Figure 2.1).

S-LAM is rare disease and prevalence was assumed to be 0.00001. I applied CEST on autosomal chromosomes and genomic control was used to adjust small inflation of our results [139].

4.3 Results

4.3.1 Evaluations of simulated samples

We evaluated the accuracy of parameter estimates using simulated data. For scenario 1, I assumed family-based samples were randomly selected, and means and standard deviations (SD) of $\hat{\beta}$ and $\widehat{h^2}$ from 2,000 replicates are given in Table 4.1. The true value of β is assumed to be 0.1253, and estimates for β by LTMH always provide a close approximation of true values. For $\widehat{h^2}$, estimates for LTMH and GCTA are similar if the prevalence is 0.1 or 0.2, although standard errors caused by estimates using LTMH are always smaller than those produced by GCTA. If prevalence is 0.05 and heritability is 0.4, bias of estimates by GCTA becomes much larger. Figure 4.2 indicates the distribution of $\widehat{h^2}$, and both methods accurately estimate high prevalence. Estimates generated by GCTA, however, are more widely distributed than those generated by LTMH, and I can conclude that LTMH provides generally superior performance.

Table 4.2 provides summaries of parameter estimates for ascertained families. According to the results, the majority of GCTA estimates are 0 and these estimates exhibit ascertainment bias. Estimates of β and h^2 by LTMH, however, are always close to true

values and these results show robustness against ascertainment bias (Table 4.2). Interestingly, standard errors resulting from estimates generated by LTMH analysis of ascertained families are small compared to those observed in the absence of ascertainment. The number of affected individuals is expected to be very small for rare diseases, but ascertainment of affected probands and familial correlations increase the number of affected individuals, which may explain the smaller standard errors observed in heritability estimates of ascertained families. Further investigation, however, is required. I also evaluated the performance of CEST in the context of hypothesis testing for scenario 1. I assumed $H_0: h^2 = 0$, and results detailing empirical sizes are given in Table 4.3. Our results indicate that LTMH analyses were slightly conservative if q = 0.05 or 0.2, but type-1 error estimates generated by this method are very close to nominal significance levels if q = 0.1. This conservative trend may indicate overestimation of variance. Table 4.3 also details the statistical power estimates. I assumed that the true h^2 is 0.2 or 0.4, and q is 0.05, 0.1 and 0.2. The statistical power estimates increase as the true heritability, prevalence, or both increase, and large empirical power estimates were obtained in regard to the larger prevalence. I also evaluated the statistical performance of the score tests for β (Table 4.4). Analyses indicate that

the score tests for β are not conservative and always preserve the nominal significance level under the null hypothesis, where $H_0: \beta = 0$. Empirical power estimates for β were assessed using 2,000 replicates at several significance levels, and these estimates increase as the prevalence, heritability, or both become larger. I also assessed empirical size estimates assuming $H_0: h^2 = 0$ for scenario 2 (Table 4.5). It was more conservative but statistical powers were improved when true h^2 is 0.2 or 0.4 than those for scenario 1.

Table 4.1 Accuracy of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{h^2}}$ from randomly selected families (scenario 1). Parameter estimates from 2,000 replicates were summarized using mean (top) and standard error (bottom). The true value of β is 0.1253. SD is standard deviation.

Heritability	Prevalence	LTMH		GCTA
		β	h^2	h^2
0.05	0.05	0.1226 (0.0223)	0.0933 (0.0971)	0.1105 (0.1303)
	0.1	0.1281 (0.0181)	0.0660 (0.0716)	0.0734 (0.0828)
	0.2	0.1277 (0.016)	0.0584 (0.0538)	0.0563 (0.0539)
0.2	0.05	0.1267 (0.0223)	0.2184 (0.1282)	0.2511 (0.1852)
	0.1	0.1239 (0.0190)	0.1950 (0.0993)	0.2111 (0.1219)
	0.2	0.1285 (0.0164)	0.2106 (0.0725)	0.2115 (0.0775)
0.4	0.05	0.1309 (0.0229)	0.4324 (0.1313)	0.5546 (0.2437)
	0.1	0.1276 (0.0225)	0.4230 (0.1315)	0.4825 (0.1377)
	0.2	0.1286 (0.0189)	0.4181 (0.0950)	0.4486 (0.085)

Figure 4.2 Boxplots for $\widehat{h^2}$ for randomly selected families (scenario 1). True heritability was 0.05 (top), 0.2 (middle), and 0.4 (bottom) and was indicated as a gray dashed line.

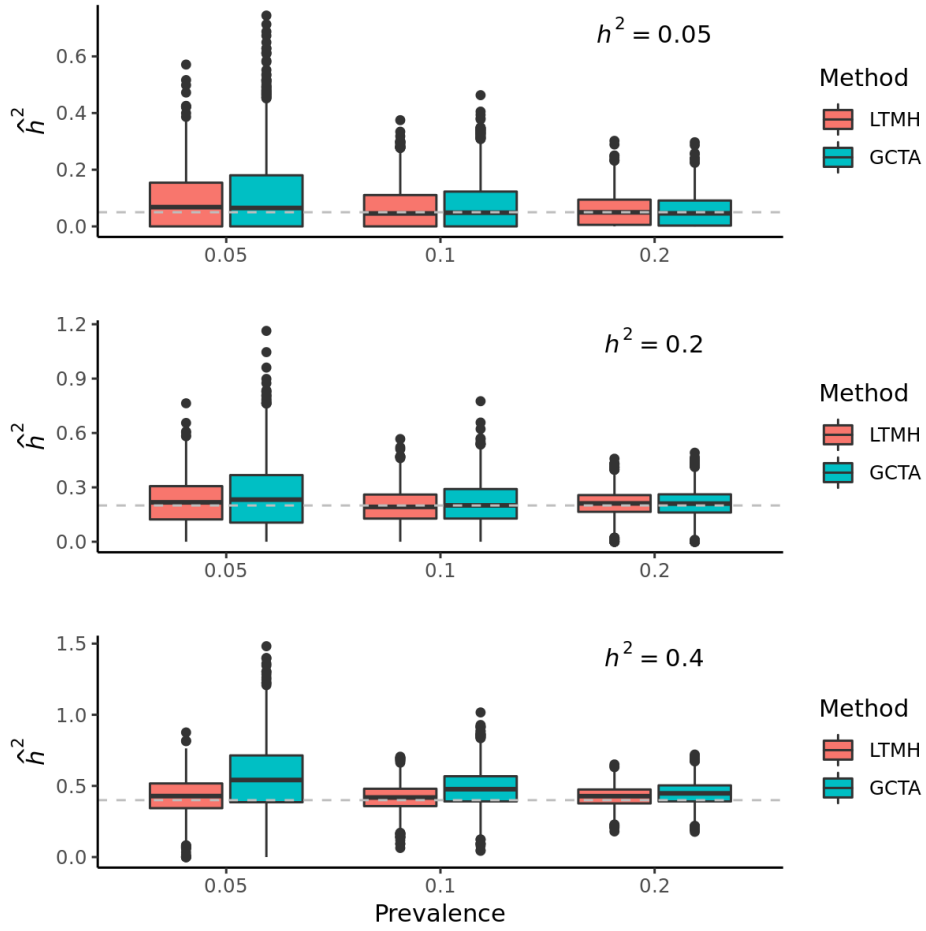


Table 4.2 Accuracy of $\widehat{\beta}$ and $\widehat{h^2}$ from ascertained families (scenario 2). Parameter estimates from 2,000 replicates were summarized using mean (top) and standard error (bottom). The true value of β is 0.1253.

Heritability	Prevalence	LTMH		GCTA
		β	h^2	h^2
0.05	0.05	0.1335 (0.0193)	0.0474 (0.0376)	1.72×10^{-6} (4.47×10^{-7})
	0.1	0.1233 (0.0181)	0.0336 (0.0339)	1.96×10^{-6} (2.01×10^{-7})
	0.2	0.1194 (0.0144)	0.0304 (0.0287)	1.83×10^{-6} (3.78×10^{-7})
0.2	0.05	0.1234 (0.0199)	0.2018 (0.0437)	1.01×10^{-6} (9.18×10^{-8})
	0.1	0.1257 (0.0135)	0.2086 (0.0342)	0 (0)
	0.2	0.1239 (0.0153)	0.1692 (0.0407)	1.01×10^{-6} (7.40×10^{-8})
0.4	0.05	0.1358 (0.0189)	0.4004 (0.0449)	0 (0)
	0.1	0.1167 (0.0144)	0.3868 (0.0339)	0 (0)
	0.2	0.1186 (0.0150)	0.4090 (0.0444)	0 (0)

Table 4.3 Type-1 error and power estimates of the proposed test for $H_0: h^2 = 0$ under scenario 1. The empirical sizes ($h^2 = 0$) and powers ($h^2 = 0.2$ and 0.4) were estimated using 2,000 replicates at three significance levels. I considered prevalence of 0.05, 0.1, and 0.2.

Heritability	Prevalence	Significance level		
		0.01	0.05	0.1
0	0.05	0.0015	0.0115	0.0285
	0.1	0.0050	0.0480	0.1020
	0.2	0.0015	0.0200	0.0505
0.2	0.05	0.0485	0.2260	0.3990
	0.1	0.3420	0.6730	0.8055
	0.2	0.6210	0.8675	0.9405
0.4	0.05	0.4575	0.8190	0.9050
	0.1	0.9395	0.9930	0.9960
	0.2	1.0000	1.0000	1.0000

Table 4.4 Type-1 error and power estimates of the proposed test for $H_0:\boldsymbol{\beta} = \mathbf{0}$ under scenario 1. The empirical sizes ($h_a^2 = 0$) and powers ($h_a^2 = 0.005$) were estimated using 2,000 replicates at three significance levels. I considered heritability of 0.2 and 0.4, and prevalence of 0.1 and 0.2.

h_a^2	Heritability	Prevalence	Significance level		
			0.01	0.05	0.1
0	0.2	0.1	0.0155	0.0661	0.1023
		0.2	0.0120	0.0560	0.0900
	0.4	0.1	0.0060	0.0480	0.0940
		0.2	0.0130	0.0580	0.1020
0.005	0.2	0.1	0.1303	0.3372	0.4713
		0.2	0.4460	0.6800	0.7980
	0.4	0.1	0.2740	0.5340	0.6640
		0.2	0.3540	0.6000	0.7180

Table 4.5 Type-1 error and power estimates of the proposed test for $H_0: h^2 = 0$ under scenario 2. The empirical sizes ($h^2 = 0$) and powers ($h^2 = 0.2$ and 0.4) were estimated using 2,000 replicates at three significance levels. I considered prevalence of 0.05, 0.1, and 0.2.

Heritability	Prevalence	Significance level		
		0.01	0.05	0.1
0	0.05	0.0000	0.0025	0.0100
	0.1	0.0005	0.0045	0.0095
	0.2	0.0000	0.0075	0.0215
0.2	0.05	0.4735	0.8110	0.9185
	0.1	0.8520	0.9660	0.9850
	0.2	0.8155	0.9540	0.9855
0.4	0.05	1.0000	1.0000	1.0000
	0.1	1.0000	1.0000	1.0000
	0.2	1.0000	1.0000	1.0000

4.3.2 Applications of LTMH and CEST to Type-2 Diabetes

To evaluate the performance of LTMH using real data, I examined the family-based samples from the T2D dataset. Table 4.6 shows the descriptive statistics [37]. There were 1,736 T2D patients (36.75%), and average age for entire samples was 48.63 years old with SD of 15.7 . The proportions of males and females are similar. All non-probands are the first-degree relatives of probands, and the familial relationships observed most often are siblings (59.22%) and offspring (32.85%).

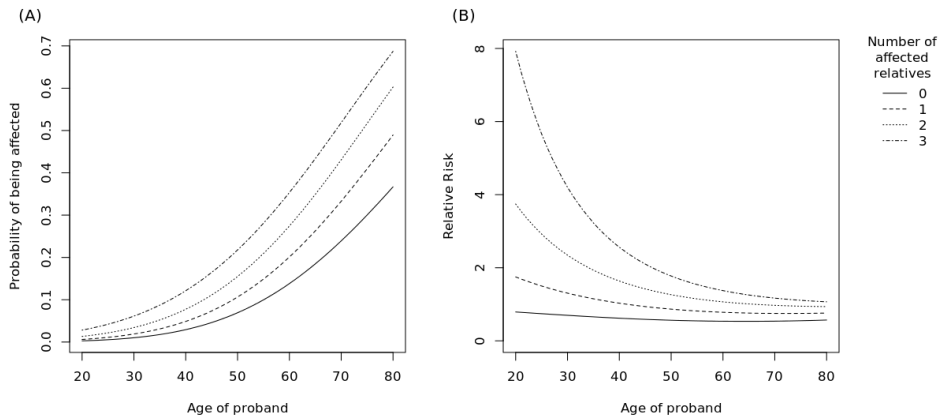
LTMH was used to examine the family-based samples derived from the T2D dataset, and heritability of T2D was estimated. Estimated heritability of T2D was 29.44%, and it was statistically significant under the significance level of 0.05 (P-value = 1.20×10^{-5}). This finding is slightly overestimated in comparison to other determinations of heritability estimates for T2D (26%) using the ACE model based on twin data [98]. This difference may be attributable to racial differences. The coefficient estimate for non-standardized age was 0.051 (0.8 for standardized age), which means that the threshold for disease is reduced by 0.051 at the liability scale if age increases by 1. The function of age is well described in Figure 4.3A, which illustrates the

probability of being affected by T2D as a function of age. Results demonstrate that the risk increases monotonically by age, reflecting the reduction effect on disease threshold. Individuals with a higher number of T2D affected relatives exhibit greater risk. In comparison to random samples, the influence of family history is greater at a young age, and determining familial risk for early-onset T2D is highly important (Figure 4.3B).

Table 4.6 Demographic characteristics of study participants. For categorical variables, the number of subjects and their proportions are provided. For continuous variables, means and standard deviations are provided. † T2D : Type 2 Diabetes.

	Proband	Non-proband
Disease status		
T2D†	648 (100%)	1,115 (26.87%)
Normal	0 (0%)	3,034 (73.13%)
Sex		
Male	308 (47.53%)	2,058 (49.6%)
Female	340 (52.47%)	2,091 (50.4%)
Age	55.44 (10.7)	47.56 (16.09)
Relationship of relatives with proband		
Parents		329 (7.93%)
Sibling		2,457 (59.22%)
Offspring		1,363 (32.85%)

Figure 4.3 Estimation of risks for T2D according to age. For a certain individual, I assume that he/she has two parents and one younger sibling, and the risk of T2D development was calculated as a function of his/her age and the number of affected family members. The X-axis indicates age of individual, and the age of his/her father and mother were assumed to be 29 years old. The younger sibling was assumed to be 3 years younger than the participant. h^2 and the coefficient of unstandardized age were set to be 0.2944 and 0.051, respectively. (A) Probability of the participant being affected according to the number of affected family members, and (B) relative risks of being affected according to the number of affected family members.



4.3.2 Applications of CEST to S-LAM disease

GWAS using CEST was performed for 549,599 SNPs. Figure 4.4 shows quantile-quantile plot and Manhattan plot of GWAS after applying genomic control. Genomic inflation factor before genomic control was 1.076. The distribution of the observed P-values met the expected P-values except two significant SNPs under the genome-wide significance level of 5×10^{-8}. CEST yielded smaller P-value for two significant SNPs rather than the result of GWAS using CLR, providing CEST is applicable to the independent samples with various strategies (Table 4.7).

Figure 4.4 Quantile-quantile (QQ) and Manhattan plots for the LAM GWAS using CEST.

a) The observed distributions of P-values for 549,599 genotyped SNPs are plotted relative to the expected (null) distribution for the CEST. b) Manhattan plot. Each dot represents the P-value of a single SNP, plotted on the genome scale at bottom. The Y-axis value is the negative logarithm of the P-value for association between each genotyped SNP and S-LAM. Two SNPs on 15q met genome-wide significance.

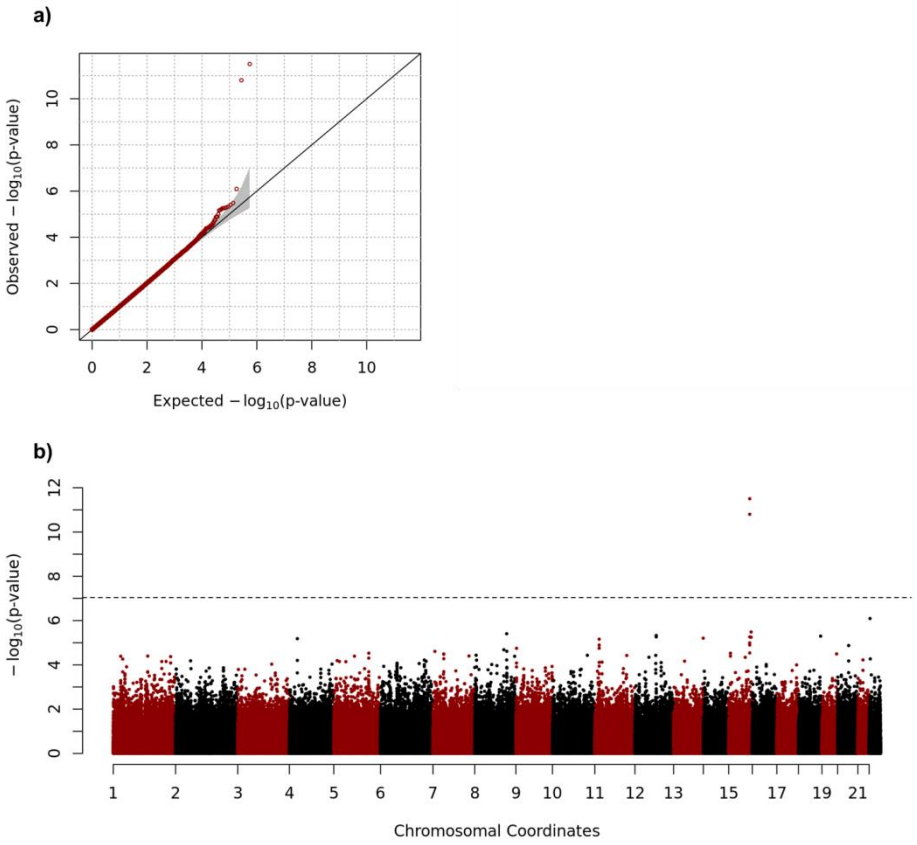


Table 4.7 Comparison results of CLR and CEST. Two significant SNPs whose P-value is less than genome-wide significance level of 5×10^{-8}.

CHR	SNP	Position	P-value	
			CLR	CEST
15	rs4544201	96167827	8.51×10^{-10}	1.581×10^{-11}
15	rs2006950	96179390	3.92×10^{-10}	3.139×10^{-12}

4.4 Discussion

In this article, I proposed a new method to estimate the heritability of a dichotomous trait based on the Liability Threshold Model for ascertained family-based samples. A simulation study demonstrated that LTMH generally provides more accurate estimates of heritabilities than does GCTA, and the differences between these methods are substantial in the context of ascertained families. To our knowledge, there is no method to effectively approach ascertained samples to estimate heritability of dichotomous traits. Additionally, I assessed the statistical performance of CEST analysis. Statistical power estimates were evaluated under various experimental conditions, and substantial power improvement was observed in the context of common diseases as opposed to that seen for rare diseases.

Despite the power improvement provided by the proposed methods, there are limitations. First, the CEST for h^2 was conservative. I found that the likelihood for h^2 is not symmetric under the null hypothesis, and this may be attributable to the misspecified weights for 0 and $\chi^2(df = 1)$ for the distribution of the CEST under H_0. Fortunately, I found that such inflation does not affect the statistical power of our analysis, but certain modifications such as bootstrapping are necessary. Second, the proposed method is the computationally

intensive when the family size, n_i, is large, and the expected computational time is proportional to $O(\max_i n_i^3)$. The most significant computational burden arises from the calculation of conditional expectation in the E-step of the EM algorithm. The computational burden can be reduced by reducing the number of iterations for the EM algorithm or by approximating the moment of the multivariate truncated normal. The former can be achieved by using EM acceleration methods which can make EM dramatically faster. These include Aitken acceleration, conjugate gradient acceleration, quasi-Newtonian acceleration, and parameter expansion acceleration [140-144]. For the latter, conditional expectation may be approximated using certain numerical algorithms such as Laplace approximation. Investigation of these techniques will be the focus of future research.

Heritability shows important utility for genetic epidemiology; however, heritability estimation of dichotomous phenotypes can be extremely complicated due to ascertainment bias. Despite several limitations, our proposed method successfully enabled heritability estimation of dichotomous traits in ascertained families, and this method may provide a promising strategy to estimate the narrow-sense heritability of various diseases. LTMH is implemented in R language,

and source codes are freely available at http://healthstat.snu.ac.kr/software/LTMH.

4.5 Appendix

4.5.1 Numerical analysis for optimization of the heritability in M-step of EM algorithm

The first derivative of $Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)$ with respect to h^2 is given by

$$\frac{\partial Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)}{\partial h^2} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i}\right)\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}_i\mathbf{A}_i^{(k)}\right)$$

$$+ \boldsymbol{\beta}^t\mathbf{X}_i^t\mathbf{C}_i\left(\mathbf{B}_i^{(k)} - \frac{1}{2}\mathbf{X}_i\boldsymbol{\beta}\right) \quad (2)$$

where $\mathbf{C}_i = \partial\boldsymbol{\Sigma}_i^{-1}/\partial h^2 = -\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i}\right)\boldsymbol{\Sigma}_i^{-1}$. Then, the objective function becomes

$$\mathcal{M}(h^2) = \sum_{i=1}^{n}\left.\frac{\partial Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(k)}\right)}{\partial h^2}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}(h^2)} = 0.$$

Similarly, I can get the first derivative of $\mathcal{M}(h^2)$ with respect to h^2 as follows,

$\mathcal{M}'(h^2)$

$$= \sum_{i=1}^{n} \Bigg[-\frac{1}{2}\mathrm{tr}\left(\mathbf{C}_i\left(\mathbf{\Phi}_i - \mathbf{I}_{n_i}\right)\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{H}_i\mathbf{A}_i^{(k)}\right)$$

$$+ \left(\mathbf{X}_i\mathbf{F}^{(k)}\right)^t\mathbf{C}_i\left(\mathbf{B}_i^{(k)} - \frac{1}{2}\mathbf{X}_i\boldsymbol{\beta}^{(k)}(h^2)\right)$$

$$+ \left(\mathbf{X}_i\boldsymbol{\beta}^{(k)}(h^2)\right)^t\mathbf{H}_i\left(\mathbf{B}_i^{(k)} - \frac{1}{2}\mathbf{X}_i\boldsymbol{\beta}^{(k)}(h^2)\right)$$

$$- \frac{1}{2}\left(\mathbf{X}_i\boldsymbol{\beta}^{(k)}(h^2)\right)^t\mathbf{C}_i\mathbf{X}_i\mathbf{F}^{(k)}\Bigg]$$

where $\mathbf{H}_i = \partial\mathbf{C}_i/\partial h^2 = -2\mathbf{\Sigma}_i^{-1}\left(\mathbf{\Phi}_i - \mathbf{I}_{n_i}\right)\mathbf{C}_i$ and

$\mathbf{F}^{(k)}$

$= \partial\boldsymbol{\beta}^{(k)}(h^2)/\partial h^2$

$$= -\left(\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{C}_i\mathbf{X}_i\right)\left(\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{\Sigma}_i^{-1}\mathbf{B}_i^{(k)}\right)$$

$$+\left(\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{C}_i\mathbf{B}_i^{(k)}\right).$$

Finally, h^2 is updated according to the following iterative steps using $\mathbf{A}_i^{(k)}$ and $\mathbf{B}_i^{(k)}$ which were calculated at the previous E-step,

$$h^2_{\text{new}} = h^2_{\text{old}} - \frac{\mathcal{M}\left(h^2_{\text{old}}\right)}{\mathcal{M}'\left(h^2_{\text{old}}\right)}.$$

4.5.2 Numerical analysis for maximizing the global lower bound

If I denote the global lower bound for the conditional log-likelihood as $\mathbb{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, then the first derivative of $\mathbb{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$, $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, is given by

$$\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \begin{pmatrix} \frac{\partial \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\beta}} + \frac{\partial l(\boldsymbol{\beta};\, \mathbf{Y}^P)}{\partial \boldsymbol{\beta}} \\ \frac{\partial \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial h^2} \end{pmatrix}.$$

Here, it should be noted that $\partial \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})/\partial \boldsymbol{\theta}$ is equivalent to the equations (1) and (2) in the Method and Appendix (A). Using the chain rule, I can easily obtain the first derivative of $l(\boldsymbol{\beta};\, \mathbf{Y}^P)$ with respect to $\boldsymbol{\beta}$ as follows,

$$\frac{\partial l(\boldsymbol{\beta};\, \mathbf{Y}^P)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{\partial l(\boldsymbol{\beta};\, Y_i^P)}{\partial \alpha_i} \frac{\partial \alpha_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right]$$

$$= \sum_{i=1}^{n} \left[\frac{(Y_i^P - \mu_i)}{\mu_i(1-\mu_i)} \phi(c - \mathbf{X}_i^P \boldsymbol{\beta})(\mathbf{X}_i^P)^t \right]$$

where $\phi(\cdot)$ is the probability density function for the standard normal. To apply Newton-Raphson algorithm for the objective function $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, I derive the first derivative of $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}^t$, $\mathbf{J}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, as follows,

$$\mathbf{J}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right) = \begin{pmatrix} \frac{\partial^2 \mathrm{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}} + \frac{\partial^2 l(\boldsymbol{\beta};\, \mathbf{Y}^P)}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}} & \frac{\partial^2 \mathrm{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)}{\partial h^2 \partial \boldsymbol{\beta}} \\ \frac{\partial^2 \mathrm{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\beta}^t \partial h^2} & \frac{\partial^2 \mathrm{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)}{\partial (h^2)^2} \end{pmatrix}$$

and each term is given by

$$\frac{\partial^2 \mathrm{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (\mathbf{X}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i), \frac{\partial^2 l(\boldsymbol{\beta};\, \mathbf{Y}^P)}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}}$$

$$= \sum_{i=1}^{n} \left[\frac{\phi(c - \mathbf{X}_i^P \boldsymbol{\beta})}{\mu_i(1-\mu_i)} \mathbf{X}_i^P \left\{ \frac{(Y_i^P - \mu_i)(2\mu_i - 1)}{\mu_i(1-\mu_i)} \phi(c \right. \right.$$

$$\left. \left. - \mathbf{X}_i^P \boldsymbol{\beta}) + (Y_i^P - \mu_i)(c - \mathbf{X}_i^P \boldsymbol{\beta}) - \phi(c - \mathbf{X}_i^P \boldsymbol{\beta}) \right\} (\mathbf{X}_i^P)^t \right],$$

$$\frac{\partial^2 \mathrm{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)}{\partial h^2 \partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(\mathbf{X}_i^t \mathbf{C}_i \mathbf{B}_i^{(k)} \right) - \sum_{i=1}^{n} (\mathbf{X}_i^t \mathbf{C}_i \mathbf{X}_i \boldsymbol{\beta}),$$

and

$$\frac{\partial^2 \mathrm{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)}{\partial (h^2)^2} = -\frac{1}{2} \mathrm{tr}\left(\mathbf{C}_i \left(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i} \right) \right) - \frac{1}{2} \mathrm{tr}\left(\mathbf{H}_i \mathbf{A}_i^{(k)} \right)$$

$$+ \boldsymbol{\beta}^t \mathbf{X}_i^t \mathbf{H}_i \left(\mathbf{B}_i^{(k)} - \frac{1}{2} \mathbf{X}_i \boldsymbol{\beta} \right).$$

With these terms, I iteratively update $\boldsymbol{\theta}$ using the following equation until convergence,

$$\boldsymbol{\theta}^{\mathrm{new}} = \boldsymbol{\theta}^{\mathrm{old}} - \mathbf{J}^{-1}\left(\boldsymbol{\theta}^{\mathrm{old}}|\boldsymbol{\theta}^{(k)}\right) \mathcal{H}\left(\boldsymbol{\theta}^{\mathrm{old}}|\boldsymbol{\theta}^{(k)}\right).$$

Chapter 5

Summary and Conclusions

Over the last few decades, genome-wide association studies (GWAS) have identified more than 69,000 variants associated with human complex traits. Rapid improvement in next-generation sequencing technology enabled us to obtain more genetic information with limited cost, but sequencing cost is still expensive. Thus, effective selection of subjects for DNA sequencing is required in order to increase statistical power of GWAS. In this thesis, I focused on incorporating family history into GWAS.

In chapter 3, I proposed a new statistical method for selecting cases and controls to improve statistical power of GWAS in sequencing DNA samples. Assuming a disease model is based on the liability threshold model, I calculated measure for selecting subjects by taking the expectation to the proband's liability conditioning his/her disease

statuses and proband's own disease status. Based on the assumption that the liabilities of related samples follow a multivariate normal distribution with variance-covariance matrix of genetic relationship matrix, I yielded the scores using moments of truncated normal distribution. Then the person who have more affected relatives might have relatively larger score than the person who have less affected relatives. In our simulation study, I considered several strategies of selecting subjects and GWAS produces largest empirical power estimates when affected subjects with large score and unaffected subjects with small score are utilized to GWAS as cases and controls, respectively. On the other hand, when affected subjects with small score and unaffected subjects with large score are used as cases and controls, GWAS worked poorly even rather than randomly selected samples. The proposed method was successively applied to T2D dataset and I found that GWAS of the proposed sample selection strategy produced lower P-value for candidate SNPs than GWAS of the randomly selected samples.

Family history has been considered as important risk factor for various complex diseases and it is relatively easy to obtain with low costs. Family history can be usually obtained via an affected families member, referring to a proband, and therefore, tends to include more

affected subjects rather than random population. Various methods to estimate heritability of binary trait have been suggested but no suitable method dealing with ascertained samples has been developed. In chapter 4, I proposed a new method to estimate heritability of binary trait on ascertained samples using conditional expectation-maximization (CEM) algorithm. In extensive simulation study, our proposed method provided accurate estimates for heritability and coefficients of covariates for both randomly selected families and ascertained families. I successfully applied the proposed model to T2D datasets consisting of ascertained families. In LAM dataset, I matched one cases with two controls based on age and top two PC scores, and performed GWAS using CEST as if matched samples are a family. In comparison to conditional logistic regression, the proposed method showed smaller P-values for two significant SNPs.

In summary, I found that a strategy of selecting cases and controls for GWAS can affect statistical power, and substantial improvement in statistical power of GWAS can be achieved by incorporating family history to selection strategy of subjects. Therefore, the proposed selection strategy seems to be cost-effective and efficient method in that I choose study participants who can most effectively detect GWAS signals based on the family history. I also proposed a

new method to estimate heritability of dichotomous phenotype for ascertained samples. Although there are some limitations, the proposed method successfully performed in both simulation study and real data analysis. Both methods in chapter 3 and 4 were implemented in R language, and source codes and manuals are freely available at websites.

Bibliography

1. Visscher, P.M., et al., *Five years of GWAS discovery.* The American Journal of Human Genetics, 2012. **90**(1): p. 7-24.
2. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.
3. Psychiatric, G.C.C.C., et al., *Genomewide association studies: history, rationale, and prospects for psychiatric disorders.* Am J Psychiatry, 2009. **166**(5): p. 540-56.
4. Livak, K.J., J. Marmaro, and J.A. Todd, *Towards fully automated genome-wide polymorphism screening.* Nature genetics, 1995. **9**(4): p. 341.
5. Scurrah, K.J., L.J. Palmer, and P.R. Burton, *Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (GLMMs) and Gibbs sampling in BUGS.* Genet Epidemiol, 2000. **19**(2): p. 127-48.
6. Zhou, W., et al., *Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies.* Nature Genetics, 2018. **50**(9): p. 1335-1341.

7. Wang, S.-B., et al., *Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology*. Scientific Reports, 2016. **6**: p. 19444.
8. Consortium, I.H., *The international HapMap project*. Nature, 2003. **426**(6968): p. 789.
9. Consortium, G.P., *A map of human genome variation from population-scale sequencing*. Nature, 2010. **467**(7319): p. 1061.
10. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data*. Nature, 2018. **562**(7726): p. 203.
11. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease*. science, 2008. **322**(5903): p. 881-888.
12. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges*. Nature reviews genetics, 2008. **9**(5): p. 356.
13. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)*. Nucleic acids research, 2016. **45**(D1): p. D896-D901.
14. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the royal statistical society. Series B (methodological), 1977: p. 1-38.

15. Gianola, D., *Heritability of polychotomous characters.* Genetics, 1979. **93**(4): p. 1051-1055.

16. Van Vleck, L., *Estimation of heritability of threshold characters.* Journal of Dairy Science, 1972. **55**(2): p. 218-225.

17. Gianola, D., *Genetic evaluation of animals for traits with categorical responses.* Journal of Animal Science, 1980. **51**(6): p. 1272-1276.

18. Stiratelli, R., N. Laird, and J.H. Ware, *Random-effects models for serial observations with binary response.* Biometrics, 1984: p. 961-971.

19. Magnussen, S. and A. Kremer, *The beta-binomial model for estimating heritabilities of binary traits.* Theoretical and applied genetics, 1995. **91**(3): p. 544-552.

20. Paul, A.K. and V. Bhatia, *Modification of beta-binomial method of estimation of heritability of stayability.* J. Ind. Soc. Agril. Statist, 2001. **54**(3): p. 385-395.

21. Papachristou, C., C. Ober, and M. Abney, *Genetic variance components estimation for binary traits using multiple related individuals.* Genetic epidemiology, 2011. **35**(5): p. 291-302.

22. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* The American Journal of Human Genetics, 2011. **88**(1): p. 76-82.

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

24. Kitaichi, M., et al., *Pulmonary lymphangioleiomyomatosis: a report of 46 patients including a clinicopathologic study of prognostic factors.* American journal of respiratory and critical care medicine, 1995. **151**(2): p. 527-533.

25. Chu, S.C., et al., *Comprehensive evaluation of 35 patients with lymphangioleiomyomatosis.* CHEST Journal, 1999. **115**(4): p. 1041-1052.

26. Urban, T., et al., *Pulmonary lymphangioleiomyomatosis: a study of 69 patients.* MEDICINE-BALTIMORE-, 1999. **78**: p. 321-337.

27. Cunha, B., et al., *Pulmonary Lymphangioleiomyomatosis on a Post-Menopausal Woman with Chronic Lymphocytic Leukaemia.* Case Reports in Clinical Medicine, 2016. **5**(03): p. 101.

28. Youssef, A.L., et al., *Lymphangioleiomyomatosis: An unusual age of diagnosis with literature review*. International Journal of Diagnostic Imaging, 2014. **1**(1): p. 17.

29. Soler-Ferrer, C., et al., *Lymphangioleiomyomatosis in a post-menopausal women*. Archivos de Bronconeumología ((English Edition)), 2010. **46**(3): p. 148-150.

30. Taylor, J.R., et al., *Lymphangioleiomyomatosis*. New England Journal of Medicine, 1990. **323**(18): p. 1254-1260.

31. Kalassian, K.G., et al., *Lymphangioleiomyomatosis: new insights*. American journal of respiratory and critical care medicine, 1997. **155**(4): p. 1183-1186.

32. Giannikou, K., et al., *Whole exome sequencing identifies TSC1/TSC2 biallelic loss as the primary and sufficient driver event for renal angiomyolipoma development*. PLoS genetics, 2016. **12**(8): p. e1006242.

33. Carsillo, T., A. Astrinidis, and E.P. Henske, *Mutations in the tuberous sclerosis complex gene TSC2 are a cause of sporadic pulmonary lymphangioleiomyomatosis*. Proceedings of the National Academy of Sciences, 2000. **97**(11): p. 6085-6090.

34. Moss, J., et al., *Prevalence and clinical characteristics of lymphangioleiomyomatosis (LAM) in patients with tuberous*

sclerosis complex. American journal of respiratory and critical care medicine, 2001. **164**(4): p. 669-671.

35. Regan, E.A., et al., *Genetic epidemiology of COPD (COPDGene) study design.* COPD: Journal of Chronic Obstructive Pulmonary Disease, 2011. **7**(1): p. 32-43.

36. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* The American Journal of Human Genetics, 2007. **81**(3): p. 559-575.

37. Song, Y.E., et al., *ONETOOL for the analysis of family-based big data*. Bioinformatics, 2018. **1**: p. 3.

38. Wigginton, J.E., D.J. Cutler, and G.R. Abecasis, *A note on exact tests of Hardy-Weinberg equilibrium*. The American Journal of Human Genetics, 2005. **76**(5): p. 887-893.

39. Raymond, M. and F. Rousset, *An exact test for population differentiation.* Evolution, 1995. **49**(6): p. 1280-1283.

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

41. Sekhon, J.S., *Multivariate and propensity score matching software with automated balance optimization: the matching package for R.* 2011.

42. Therneau, T.M. and T. Lumley, *Package 'survival'.* R package version, 2017: p. 2.41-3.

43. Lee, S., et al., *Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.* The American Journal of Human Genetics, 2012. **91**(2): p. 224-237.

44. Consortium, H.R., *A reference panel of 64,976 haplotypes for genotype imputation.* Nature genetics, 2016. **48**(10): p. 1279-1283.

45. Delaneau, O., J. Marchini, and G.P. Consortium, *Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel.* Nature communications, 2014. **5**: p. 3934.

46. Durbin, R., *Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT).* Bioinformatics, 2014. **30**(9): p. 1266-1272.

47. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nature reviews. Genetics, 2010. **11**(7): p. 499.

48. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2004. **21**(2): p. 263-265.

49. Farh, K.K.-H., et al., *Genetic and epigenetic fine mapping of causal autoimmune disease variants.* Nature, 2015. **518**(7539): p. 337-343.

50. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-380.

51. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome biology, 2013. **14**(4): p. R36.

52. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC bioinformatics, 2011. **12**(1): p. 323.

53. Network, C.G.A.R., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061.

54. Lonsdale, J., et al., *The genotype-tissue expression (GTEx) project.* Nature genetics, 2013. **45**(6): p. 580-585.

55. Bongaarts, A., et al., *Subependymal giant cell astrocytomas in Tuberous Sclerosis Complex have consistent TSC1/TSC2 biallelic inactivation, and no BRAF mutations.* Oncotarget, 2017. **8**(56): p. 95516.

56. Poirier, J.G., et al., *Resampling to Address the Winner's Curse in Genetic Association Analysis of Time to Event*. Genetic epidemiology, 2015. **39**(7): p. 518-528.

57. Grubert, F., et al., *Genetic control of chromatin states in humans involves local and distal chromosomal interactions.* Cell, 2015. **162**(5): p. 1051-1065.

58. Qiu, Y., et al., *Isolation, characterization, and chromosomal localization of mouse and human COUP-TF I and II genes.* Genomics, 1995. **29**(1): p. 240-6.

59. Rada-Iglesias, A., et al., *Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest.* Cell stem cell, 2012. **11**(5): p. 633-648.

60. Julian, L.M., et al., *Human Pluripotent Stem Cell–Derived TSC2-Haploinsufficient Smooth Muscle Cells Recapitulate Features of Lymphangioleiomyomatosis.* Cancer research, 2017. **77**(20): p. 5491-5502.

61. Qin, J., et al., *COUP-TFII regulates tumor growth and metastasis by modulating tumor angiogenesis.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(8): p. 3687-3692.

62. Xu, M.F., et al., *The role of the orphan nuclear receptor COUP-TFII in tumorigenesis.* Acta Pharmacologica Sinica, 2015. **36**(1): p. 32-36.

63. Juvet, S.C., D. Hwang, and G.P. Downey, *Rare lung diseases I--Lymphangioleiomyomatosis.* Can Respir J, 2006. **13**(7): p. 375-80.

64. McCormack, F.X., et al., *Official American Thoracic Society/Japanese Respiratory Society Clinical Practice Guidelines: Lymphangioleiomyomatosis Diagnosis and Management*. American Journal of Respiratory and Critical Care Medicine, 2016. **194**(6): p. 748-761.

65. Riggs, K.A., et al., *Decreased chicken ovalbumin upstream promoter transcription factor II expression in tamoxifen-resistant breast cancer cells.* Cancer Research, 2006. **66**(20): p. 10188-10198.

66. Glasgow, C.G., et al., *Lymphatic involvement in lymphangioleiomyomatosis.* Lymphatic Continuum Revisited, 2008. **1131**: p. 206-214.

67. Seyama, K., K. Mitani, and T. Kumasaka, *Lymphangioleiomyoma Cells and Lymphatic Endothelial Cells Expression of VEGFR-3 in Lymphangioleiomyoma Cell*

Clusters. American Journal of Pathology, 2010. **176**(4): p. 2051-2052.

68. Young, L.R., et al., *Serum VEGF-D concentration as a biomarker of lymphangioleiomyomatosis severity and treatment response: a prospective analysis of the Multicenter International Lymphangioleiomyomatosis Efficacy of Sirolimus (MILES) trial.* Lancet Respiratory Medicine, 2013. **1**(6): p. 445-452.

69. Srinivasan, R.S., et al., *The nuclear hormone receptor Coup-TFII is required for the initiation and early maintenance of Prox1 expression in lymphatic endothelial cells.* Genes & development, 2010. **24**(7): p. 696-707.

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

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

72. Sboner, A., et al., *The real cost of sequencing: higher than you think!* Genome biology, 2011. **12**(8): p. 125.

73. Moore, G.E., *Cramming more components onto integrated circuits.* Proceedings of the IEEE, 1998. **86**(1): p. 82-85.

74. Maher, B., *Personal genomes: The case of the missing heritability.* Nature News, 2008. **456**(7218): p. 18-21.

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

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

77. Antoniou, A.C. and D.F. Easton, *Polygenic inheritance of breast cancer: implications for design of association studies.* Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 2003. **25**(3): p. 190-202.

78. Howson, J.M., et al., *Comparison of population-and family-based methods for genetic association analysis in the presence of interacting loci.* Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 2005. **29**(1): p. 51-67.

79. Li, M., M. Boehnke, and G.R. Abecasis, *Efficient study designs for test of genetic association using sibship data and unrelated cases and controls.* The American Journal of Human Genetics, 2006. **78**(5): p. 778-792.

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

81. Kim, W., et al., *Selecting cases and controls for DNA sequencing studies using family histories of disease.* Statistics in medicine, 2017. **36**(13): p. 2081-2099.

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

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

84. Stroup, W.W., *Generalized Linear Mixed Models: Modern Concepts.* Methods and Applications Bosa Roca: CRC Press, 2012.

85. Bliss, C.I., *The method of probits—a correction.* Science, 1934. **79**(2053): p. 409-410.

86. Wilhelm, S. and B. Manjunath, *tmvtnorm: A package for the truncated multivariate normal distribution.* sigma, 2010. **2**(2).

87. Vazquez, A., et al., *An R package for fitting generalized linear mixed models in animal breeding 1.* Journal of animal science, 2010. **88**(2): p. 497-504.

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

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

90. Barnett, I.J., S. Lee, and X. Lin, *Detecting rare variant effects using extreme phenotype sampling in sequencing association studies.* Genetic epidemiology, 2013. **37**(2): p. 142-151.

91. Nebert, D.W., *Extreme discordant phenotype methodology: an intuitive approach to clinical pharmacogenetics.* European journal of pharmacology, 2000. **410**(2-3): p. 107-120.

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

93. Risch, N.J. and H. Zhang, *Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations.* American Journal of Human Genetics, 1996. **58**(4): p. 836.

94. Gelman, A. and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. 2006: Cambridge university press.

95. Cochran, W.G., *Some Methods for Strengthening the Common χ^2 Tests*. Biometrics, 1954. **10**(4): p. 417-451.

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

97. Kim, D.J., *The Epidemiology of Diabetes in Korea*. Diabetes Metab J, 2011. **35**(4): p. 303-308.

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

99. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies*. Nature Genetics, 2006. **38**: p. 904.

100. Weiss, N.A., *A course in probability*. 2006: Addison-Wesley.

101. Risch, N. and J. Teng, *The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling*. Genome Res, 1998. **8**(12): p. 1273-88.

102. Wang, J. and S. Shete, *Analysis of secondary phenotype involving the interactive effect of the secondary phenotype and genetic variants on the primary disease.* Ann Hum Genet, 2012. **76**(6): p. 484-99.

103. van der Sluis, S., D. Posthuma, and C.V. Dolan, *TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies.* PLoS Genet, 2013. **9**(1): p. e1003235.

104. Li, H. and M.H. Gail, *Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies.* Hum Hered, 2012. **73**(3): p. 159-73.

105. Schifano, E.D., et al., *Genome-wide association analysis for multiple continuous secondary phenotypes.* Am J Hum Genet, 2013. **92**(5): p. 744-59.

106. O'Reilly, P.F., et al., *MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS.* PLoS One, 2012. **7**(5): p. e34861.

107. Lee, S.H. and J.H. van der Werf, *MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information.* Bioinformatics, 2016. **32**(9): p. 1420-2.

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

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

110. McCulloch, C.E. and J.M. Neuhaus, *Generalized Linear Mixed Models.* Wiley StatsRef: Statistics Reference Online, 2014.

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

112. Fitzmaurice, G., et al., *Longitudinal data analysis*. 2008: CRC Press.

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

114. Mi, X., T. Miwa, and T. Hothorn, *mvtnorm: New numerical algorithm for multivariate normal probabilities.* The R Journal, 2009. **1**(1): p. 37-39.

115. Visscher, P.M., W.G. Hill, and N.R. Wray, *Heritability in the genomics era—concepts and misconceptions.* Nature reviews genetics, 2008. **9**(4): p. 255.

116. Fedko, I.O., et al., *Estimation of genetic relationships between individuals across cohorts and platforms: application to childhood height*. Behavior genetics, 2015. **45**(5): p. 514-528.

117. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height*. Nature genetics, 2010. **42**(7): p. 565.

118. Vattikuti, S., J. Guo, and C.C. Chow, *Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits*. PLoS genetics, 2012. **8**(3): p. e1002637.

119. Dudbridge, F., *Power and predictive accuracy of polygenic risk scores*. PLoS genetics, 2013. **9**(3): p. e1003348.

120. Burton, P.R., et al., *Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling*. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 1999. **17**(2): p. 118-140.

121. Dempster, E.R. and I.M. Lerner, *Heritability of threshold characters*. Genetics, 1950. **35**(2): p. 212.

122. Hoeschele, I. and B. Tier, *Estimation of variance components of threshold characters by marginal posterior modes and means*

via Gibbs sampling. Genetics Selection Evolution, 1995. **27**(6): p. 519.

123. Sawyer, S., *Maximum likelihood estimators for incorrect models, with an application to ascertainment bias for continuous characters*. Theoretical Population Biology, 1990. **38**(3): p. 351-366.

124. Park, S., et al., *Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families*. BMC medical genetics, 2015. **16**(1): p. 62.

125. Jebara, T. and A. Pentland. *Maximum conditional likelihood via bound maximization and the CEM algorithm*. in *Advances in neural information processing systems*. 1999.

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

127. Abney, M., M.S. McPeek, and C. Ober, *Estimation of variance components of quantitative traits in inbred populations*. The American Journal of Human Genetics, 2000. **66**(2): p. 629-650.

128. Jacquard, A., *The genetic structure of populations*. Vol. 5. 2012: Springer Science & Business Media.

129. Atkinson, K.E., *An introduction to numerical analysis*. 2008: John Wiley & Sons.

130. Bertsekas, D.P., *Constrained optimization and Lagrange multiplier methods*. 2014: Academic press.

131. Kuhn, H.W. and A.W. Tucker, *Nonlinear programming*, in *Traces and emergence of nonlinear programming*. 2014, Springer. p. 247-258.

132. Neal, R.M. and G.E. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, in *Learning in graphical models*. 1998, Springer. p. 355-368.

133. Fisher, R.A. *Theory of statistical estimation*. in *Mathematical Proceedings of the Cambridge Philosophical Society*. 1925. Cambridge University Press.

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

135. Rao, C.R. *Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation*. in *Mathematical Proceedings of the Cambridge Philosophical Society*. 1948. Cambridge University Press.

136. Scott, W.A., *Maximum likelihood estimation using the empirical fisher information matrix*. Journal of Statistical Computation and Simulation, 2002. **72**(8): p. 599-611.

137. Organization, W.H., *Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation*. 2006.

138. Ko, S.-H., et al., *Past and Current Status of Adult Type 2 Diabetes Mellitus Management in Korea: A National Health Insurance Service Database Analysis*. Diabetes & metabolism journal, 2018. **42**(2): p. 93-100.

139. Devlin, B. and K. Roeder, *Genomic control for association studies*. Biometrics, 1999. **55**(4): p. 997-1004.

140. Laird, N., N. Lange, and D. Stram, *Maximum likelihood computations with repeated measures: application of the EM algorithm*. Journal of the American Statistical Association, 1987. **82**(397): p. 97-105.

141. Jamshidian, M. and R.I. Jennrich, *Conjugate gradient acceleration of the EM algorithm*. Journal of the American Statistical Association, 1993. **88**(421): p. 221-228.

142. Lange, K., *A quasi-Newton acceleration of the EM algorithm*. Statistica sinica, 1995: p. 1-18.

143. Lange, K., *A gradient algorithm locally equivalent to the EM algorithm*. Journal of the Royal Statistical Society. Series B (Methodological), 1995: p. 425-437.

144. Liu, C., D.B. Rubin, and Y.N. Wu, *Parameter expansion to accelerate EM: the PX-EM algorithm*. Biometrika, 1998. **85**(4): p. 755-770.

국 문 초 록

최근 유전자 시퀀싱 기술의 발전은 질병을 가진 인간의 유전정보를 대량으로 얻어내는 것을 가능하게 하였으며 이를 통하여 인간의 질병과 유전적 변이 사이의 연관성을 밝혀낼 수 있었다. 그러나 시퀀싱 기술의 발전으로 비용이 현저히 낮아졌다고 할지라도 유전정보를 얻는데 필요한 비용은 결코 저렴하지 않으며, 제한된 비용에서 최대의 효율을 끌어낼 수 있는 분석 대상을 선별하는 과정은 매우 중요하다. 한편, 이분형 표현형의 유전율을 추정하는 수많은 방법이 제안되었지만 연속형 표현형의 유전율 추정과는 달리 계산적으로 또 통계적으로 매우 복잡하여 제한적으로 이용되곤 하였다.

이에 본 논문에서는, 전장유전체연관성분석의 통계적 검정력을 향상시키기 위하여 유전자 시퀀싱을 함에 있어 가족력을 바탕으로 사례군과 대조군을 선별하는 새로운 통계적 방법을 개발하였다. 질병 모형은 관측되지 않은 연속형 변수에 의해 결정된다고 가정하는데, 이 연속형 변수가 질병

고유의 한계점보다 큰 사람은 질병을 얻게 된다. 이 연속형 변수는 책임점수(Liability) 라고 일컫고 이 질병 모형을 책임한계모형(Liability threshold model)이라고 부른다. 이 질병 모형을 바탕으로 본 연구의 방법은 다음의 두 단계로 이루어져 있다. 첫째로, 각 가족 별로 가족들의 질병력이 주어졌을 때의 책임점수의 조건부평균을 계산한다. 그 다음으로 이렇게 구해진 조건부평균을 바탕으로 사례군과 대조군을 선별한다. 모의실험을 통하여 전장유전체연관성분석의 통계적 검정력은 어떻게 사례군과 대조군을 선별하는지에 따라서 중대한 영향을 받고, 조건부평균이 큰 질병군을 사례군으로, 작은 정상군을 대조군으로 선별하였을 때 가장 높은 것을 확인하였다. 이 방법은 제 2 형 당뇨의 유전체 연관성 분석에 적용되었고, 무작위로 분석대상을 추출하였을 때와 결과와 비교하였을 때, 훨씬 더 향상된 것을 확인할 수 있었다.

이 방법과 더불어, 나는 이분형 표현형의 유전율 추정방법을 개발하였다. 이 방법은 가족력을 바탕으로 추정이 되고, 가계도의 구조에 구애 받지 않는다. 특히 이 방법은 무작위로 선별된 가족에 대한 추정 뿐 아니라, proband 의

질병력으로 인하여 가족이 분석에 참여하게 된 경우에 대한 추정도 가능하다는 장점을 가지고 있다. 다양한 모의실험을 통하여 이 방법의 정확성을 평가하였으며, 기 개발된 연구의 결과와 비교를 통하여 추정치의 정확성의 향상을 확인할 수 있었다. 또한 제 2 형 당뇨의 한국인 가계도 데이터에 본 방법을 적용하여 유전율을 평가하였다.

키워드 : 전장유전체연관성분석, 질병가족력, 위험도예측, 유전율, 책임한계모형, 확인오차

학번:2015-30118