Statistical Methods and Privacy Preserving Protocols for Combining Genetic Data with Electronic Health Records

by

Xutong Zhao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2019

Doctoral Committee:

Professor Gonçalo Abecasis, Chair
Associate Professor Hyun Min Kang
Associate Professor Seunggeun Shawn Lee
Professor Patricia A. Peyser

Xutong Zhao
xtzhao@umich.edu
ORCID id: 0000-0003-3179-7369

To my parents and Zhengling

ACKNOWLEDGEMENTS

I feel so grateful that I have had the opportunity over the past four years working on exciting and challenging problems of statistical genetics, and turning them into my Ph.D. dissertation. It is one of the most enjoyable and rewarding Ph.D. experiences that anybody could hope for. I could not imagine having better support from my advisor, committee members, colleagues, friends and family.

First of all, I would like to express my sincere appreciation to my advisor Dr. Gonçalo Abecasis for his tremendous help with my Ph.D. studies and related research, and for his patience and encouragement. He guided me to the world of statistical genetics, and taught me how to learn, to think and to question with his profound insight and extensive experience. He inspires me for the value the seeking of the simplest but most useful solutions and then persisting. During the past years, he has provided me with numerous research opportunities and guidance to build a solid foundation in genetics and biostatistics. His technical and editorial advice was essential to the completion of this dissertation.

I would also like to thank my committee members: Dr. Hyun Min Kang and Dr. Seunggeun Lee for their continued support and guidance over the past years, and Dr. Patricia A. Peyser for her advice and assistance with this thesis. Those invaluable comments have greatly improved this dissertation.

I especially want to thank Sayantan Das for his mentorship during the initial years of my Ph.D. study, Daniel Taliun and Sarah Gagliano Taliun for their tremendous support on analyzing Trans-Omics for Precision Medicine data, Ani Manichaikul and Michael Cho for sharing their domain expertise in the whole genome sequencing analysis of chronic obstructive pulmonary disease and leading me to this fascinating field. I am also thankful to my fellow colleagues in the Abecasis' research group, Alan Kwong, Anita Pandit and Gregory Zajac for their great exchange of scientific ideas.

I am very grateful to all my friends, who always support me and share my laughs and tears. Special appreciation alphabetically goes to Cui Guo, Peng Liao, Lu Tang, Lu Xia, Nalingna Yuan, Minling Zhang and Jinyang Zheng.

In addition, I would like to express my appreciation to Kirsten Herold, who has helped me on my scientific writing tremendously.

Last but not least, nothing would be achieved without the unconditional love and dedication of my family throughout my Ph.D. studies and my life, especially my parents and my husband Zhengling Qi. It is their support and understanding that help me get through hard times and keep me brave to face challenges.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AF: Allele Frequency

AMD: Age-related Macular Degeneration

CPU: Central Processing Unit

EHR: Electronic Health Records

eMERGE: Electronic Medical Records and Genomics

FDR: False Discovery Rate

FHE: Fully Homomorphic Encryption

FV: Fan and Vercauteren Encryption Scheme

GB: Gigabyte

gnomAD: Genome Aggregation Dataset

GWAS: Genome-wide Association Studies

HGDP: Human Genome Diversity Project

HWE: Hardy-Weinberg Equilibrium

ICD-9-CM: International Classification of Disease

KING: Kinship based Inference for Genome-wide Association Studies

LD: Linkage Disequilibrium

MAF: Minor Allele Frequency

MB: Megabytes

MGI: Michigan Genomics Initiative

MLE: Maximum Likelihood Estimates

PheWAS: Phenome-wide Association Studies

RAF: Risk Allele Frequency

RMSE: Root Mean Square Error

ROC: Receiver Operating Characteristic

SWHE: Somewhat Homomorphic Encryption Scheme

T2D: Type II Diabetes

TB: Terabytes

TOPMed: Trans-Omics for Precision Medicine

TPR: True Positive Rate

UM: The University of Michigan

ABSTRACT

In recent years, electronic health records (EHR) have been combined with genetic data to uncover disease biology and accelerate generation of hypotheses for drug development and treatment strategies. The goal of this dissertation is to develop novel statistical models that can address the challenges of analyzing 'imperfect' EHR data and to propose privacy-preserving methods that enable sensitive individual-level data sharing across EHR studies and other large genetic studies.

In Chapter II, we propose a statistical method to address misclassified clinical outcomes, a common challenge in EHR data. One essential step of EHR-based genome-wide association studies is constructing a cohort of cases and controls for a specific disease from billing codes and other clinical or administrative data. Nearly always, a perfect strategy for deriving disease phenotypes from billing codes is not available, resulting in some incorrect case/control labels. Here, we propose a method to estimate the misclassification of case/control status by examining genotype information of dozens of disease associated loci. Through simulation and application to the Michigan Genomics Initiative data, we demonstrate that the method enables the evaluation of new EHR-based phenotype definition schemes and provides accurate estimates of disease association measures when phenotypes are misclassified.

In Chapters III and IV, we focus on identifying overlapping samples between studies, a common challenge when aggregating information across datasets. We particularly focus on identifying duplicate or related samples when sharing the underlying individual level genetic data is restricted. We propose methods that do not require disclosure of individual identities but

that can still identify genetic relatives across datasets. In Chapter III, we show that by grouping genotypes into segments and calculating summary statistics within each segment, we are able to obscure and encode individual-level genetic information. Relatives can be inferred with the coded genotypes using a likelihood model. Simulation and application to the Trans-Omics for Precision Medicine (TOPMed) program data demonstrate the utility and security of the method. In Chapter IV, we extend the method further, with a strategy that guarantees stronger encryption and is expected to work across heterogeneous populations. This secure protocol can infer genetic relatives among people of diverse ethnic backgrounds. The method works by combining a cryptographic technique, homomorphic encryption, with the robust relationship inference method previously described by Manichaikul et al (2010). Through simulations, we show that our method's performance is identical to that of implementations that use the original unencrypted genotypes. Our protocol scales well in computing time and is protected from several possible attacks. The secure protocol was again applied to TOPMed dataset. Securely identifying related samples will facilitate combination of results across datasets when there are restrictions to sharing the underlying individual level data.

In conclusion, the methods developed here well enhance use of EHR data and genome data to improve accuracy of case/control status as well as decrease inclusion of relatives across studies when desired.

CHAPTER I

Introduction

1.1 Combining EHR with genetic studies

Combining the clinical data in electronic health records (EHR) with genetic data provides us a chance to accelerate the pace of genomic discovery on thousands of traits. The idea of combining DNA repositories with EHR raises the possibility that the EHR can be used in genomic research to replicate or discover genotype–phenotype associations (Denny et al., 2013; Jensen et al., 2012; Bush et al., 2016). Efforts have been devoted to link genetic data to EHR in many research programs. For example, the Electronic Medical Records and Genomics (eMERGE) Network has been funded by the National Institutes of Health (NIH) since 2007 (McCarty et al., 2011). The University of Michigan (UM) established the Michigan Genomics Initiative, which collects genotype data and EHR of patients undergoing surgery in the UM hospital. Previously, some small-scale EHR studies have demonstrated that EHR-based genome-wide association studies (GWAS) or phenome-wide association studies (PheWAS) have the ability to replicate genotype-phenotype associations as well as to uncover novel associations (Denny et al., 2013; Ritchie et al., 2013). Large cohorts can be gathered quickly and inexpensively from EHR. This advantage, as well as the reduced genotyping costs, has promoted the establishment of large biobanks, like the UK Biobank (Sudlow et al., 2015; Bycroft et al., 2018). Combining EHR-based phenotypes with the genotype data in the UK Biobank allows us to investigate associations of half a million

samples with thousands of diseases (Zhou et al., 2018; Nielsen et al., 2018; Wolford et al., 2018).

In addition to being used for GWAS and PheWAS, EHR have also been linked to sequencing data, for example, the exome sequencing project of UK Biobank samples funded by Regeneron Pharmaceuticals (UK Biobank, 2018).These sequencing data of large cohorts can be used to establish the reference dataset that is the critical resource for functional interpretation of putative disease-causing variants. Multiple similar reference resources, like dbSNP and ClinVar, or custom browsers like gnomAD and BRAVO, are becoming publicly available (Sherry et al., 2001; Landrum et al., 2017; Taliun et al., 2019; Karczewski et al., 2019). Aggregated information from multiple reference resources will help researchers make inference more efficiently and comprehensively. However, while the summaries like EHR-based GWAS results and allele frequencies (AFs) are often shared between reference datasets, studies are usually prohibited from sharing their underlying individual-level data with each other. In order to aggregate information, a vital step, then, is to infer the overlapping samples between studies. Due to the data sharing barriers, identifying genetic relatives between studies can be challenging.

Overall, incorporating genetic information into EHR brings both opportunities and challenges in many subject areas, including statistics, biology, medical science and computer science, and often requires interdisciplinary knowledge to adequately understand the problems and devise appropriate solutions. For statisticians, valid statistical analysis for large-scale EHR is one of the most important concerns. In this dissertation, we address challenging problems in EHR studies related to data misspecification as well as data privacy. Reliable and efficient solutions to some of the most important problems related to EHR studies are proposed, solutions which are general and therefore applicable to other genetic studies.

1.2 Challenges of analyzing EHR data

EHR data has several problems that may bias results if conventional statistical methods are applied. Moreover, conventional methods may not even be useable for certain purposes. In this section, we describe several problems that should be considered when dealing with EHR data, and for which we will propose methodological solutions in the following chapters.

1.2.1 Misclassification

One essential step of EHR-based GWAS is constructing a cohort of cases and controls for a specific disease, using EHR billing codes. However, a perfect rule of pooling redundant billing codes to phenotype codes is often lacking, leading to incorrect case/control outcome labels. In addition, the errors in the billing codes themselves that occur in every step of these codes' assignment also lead to the misclassification (O'malley et al.,2005). Ignoring the misclassified phenotypes can bias the association results and mislead drug and treatment research (Neuhaus, 1999; Copeland et al, 1977). For example, when we conducted preliminary GWAS of ~1500 diseases using Michigan Genomics Initiative (MGI) data, which is a collaborative study at the University of Michigan pairing patients' EHR and genetics information, we detected potential misclassification of several phenotypes like type II diabetes (T2D).

To correct the estimated effect size and increase the power of the association test in EHR GWAS when misclassification is present, several methods have been developed. Magder and Hughes (1997) proposed an unsupervised algorithm called iteratively reweighted least square algorithm to estimate misclassification rate. This method can estimate the misclassification without the identification of gold standard samples. However, due to the flat surface of the likelihood, this method cannot guarantee the convergence to the estimation that maximizes the likelihood in practice (Hong et al, 2019). Sinnott et al. (2014) also proposed a method that does not require

knowledge of the gold standard sample. Algorithms based on this method can calculate a sample-specific probability of having the disease. This method was shown to improve the test power and odds ratio estimation. However, it has limited generalizability, and the algorithms have only been developed for certain diseases like rheumatoid arthritis and Crohn's disease (Liao et al., 2010; Carroll et al., 2012; Ananthakrishnan et al., 2013).

In addition to these unsupervised methods, some supervised or semi-supervised methods have been developed (Duffy et al., 2004; Gordon et al, 2004; McDavid et al., 2013; Hong et al, 2019). These methods were shown to estimate misclassification with relatively higher accuracy and have wider generalizability compared with the unsupervised methods. However, they require identifying a set of gold standard samples with correctly specified case and control labels. Generating gold standard samples from EHR data, even a small set, requires cumbersome record review by doctors and specialists.

1.2.2 Security

Communication between different EHR studies and other sequencing studies may allow us to aggregate information of more data and conduct more powerful analyses. Inferring genetic relatives between studies is one critical step to achieve this goal. Ignoring the closely related samples will bias the aggregated information result in inaccurate interpretation of downstream analyses. For example, when the overlapping samples are enriched for a rare variants, AF of this variant in the joint population will be overestimated if we do not consider the overlapping information. In addition, for the meta-analysis of multiple GWAS, ignoring the overlaps among studies can lead to inflated type I error and false signals. However, studies are often prohibited from sharing individual-level data with each other due to privacy issues.

Homer et al. (2008) showed that given an individual's DNA information we can easily discover if he/she is involved in a GWAS. In general, various strategies and technologies have been developed to address genetic privacy problem in different areas of genetic studies. One widely accepted strategy is the dbGAP access control (Mailman et al., 2007), which protects genetic data by placing it in a secure location that is only accessible to people having the permission. Another strategy is data anonymization. Differential privacy is a typical method based on this idea. It adds reasonable noise to the summary statistic before its release. Several studies have shown that the individual information will not be revealed in the released summary statistics using the differential privacy technique (Uhlerop et al., 2013; Yu et al., 2014). Moreover, many privacy-preserving methods have been developed based on modern cryptographic solutions. For example, homomorphic encryption allows people to predict their disease susceptibility on the cloud using their encrypted genotype so that they do not need to disclose their true genetic data (Ayday et al., 2013). Although potential risks of sharing genetic data as well as techniques to protect data privacy have been widely developed, a state-of-the-art solution to deal with this specific problem, secure relationship inference, is still lacking.

1.3 Summary of objectives

With a focus on these challenges, this dissertation will propose methodologies to achieve the following analytical objectives:

1) Build a model that can estimate misclassified cases/controls in electronic health records;

2) Develop protocols allowing relationship inference without disclosing individual-level genetic data.

These two objectives are addressed in the proposed methods in Chapter II (Objective 1) and Chapters III and Chapter IV (Objective 2). More details on the background, pertinent literature, motivation and methodology development can be found in the introductions of each chapter.

CHAPTER II

Modeling Misclassified Phenotypes in Electronic Health Records Using Genotype Information

2.1 Introduction

Genome-wide association studies (GWAS) have successfully uncovered relationships between genetic factors with thousands of human traits; many of these associations between phenotypes and genetic variants require further replication. Since cohorts can be gathered efficiently from electronic health records (EHR), by combining genetic data with EHR, researchers have been able to accelerate the replication or discovery of genotype–phenotype associations (Jensen et al., 2012; Denny et al., 2013; Bush et al., 2016; Zhou et al., 2018; Nielsen et al., 2018).

To perform an EHR-based GWAS, one starts by constructing a collection of cases and controls for a specific phenotype from the EHR data. A traditional approach for phenotyping is manual chart review, which has been successfully applied to EHR data (Wilke et al, 2005). However, this method is time-consuming and cumbersome and cannot generate large cohorts easily. The billing codes, International Classification of Disease (ICD-9-CM) codes, in EHR are more commonly used to generate cases and controls (Denny et al., 2010; Denny et al., 2013; Ritchie et al., 2013). ICD-9-CM codes contain lists of codes corresponding to diagnoses and procedures recorded in conjunction with hospital care. ICD-9-CM codes are usually not necessarily direct surrogates for a phenotype as several ICD-9-CM codes may describe the same disease. For

examples, ICD code 250.00, 250.02 and eight more codes all denote to type II diabetes. Moreover, defining the control samples for a certain disease is not straightforward. Therefore, well-defined electronic phenotyping algorithms using ICD-9-CM codes, e.g. PheWAS, have been developed to dichotomize phenotypes (Denny et al., 2013). By pooling redundant billing codes to PheWAS codes, we can increase the sample size of cases as well as define a reliable set of controls.

However, even with these phenotyping algorithms, phenotyping remains challenging. Due to its billing purpose, EHR can be too flawed for scientific research. PheWAS phenotypes defined using ICD-9-CM codes usually have imperfect sensitivity, because of inherent variations in the coding scheme itself and variation in how healthcare providers assign the codes to patients (O'malley et al., 2005). Such inaccuracies typically affect the GWAS/PheWAS results by biasing results toward the null hypothesis (Neuhaus, 1999; Copeland et al, 1977). In fact, the effect sizes for EHR-based associations were typically closer to zero than those found in other large-scale GWAS. Even though this phenomenon may be due to the "winner's curse," such that the GWAS in which the association was first discovered often overestimates the true effect size, it may also be evidence of phenotype misclassification in EHR-defined case/control status (Bazerman and Samuelson, 1983; Lohmueller et al., 2003).

We detected such potential misclassification of phenotypes when analyzing Michigan Genomics Initiative (MGI) data, which is a collaborative study of the University of Michigan, pairing patients' EHR and genetics information to gain novel biomedical insights. Around 1500 phenotypes were constructed from ICD-9-CM codes through the PheWAS R package(Carroll et al., 2014). Both GWAS and PheWAS analysis were conducted on the enriched information on those traits using over 7.7M common genetic variants. The association results are displayed in

the PheWeb browser, http://pheweb.sph.umich.edu/. While several well-known genetic associations were successfully replicated, we observed thoroughly attenuated effect sizes in some traits, which may result from misclassification of the phenotypes, like age-related macular degeneration (AMD) and type II diabetes.

As described in Chapter I, both unsupervised and supervised methods have been developed to estimate the misclassification rate and correct the estimated effect size to increase the power of the association test in EHR GWAS (Magder and Hughes 1997; Duffy et al., 2004; Gordon et al, 2004; Liao et al., 2010; Carroll et al., 2012; Ananthakrishnan et al., 2013; McDavid et al., 2013; Sinnott et al. 2014; Hong et al, 2019). The unsupervised methods have disadvantages that either cannot guarantee the convergence to the maximum likelihood estimates in practice or have limited generality. While having better performance of estimating misclassification rates, the supervised methods need a time-consuming set of gold standard samples to train the model.

In this chapter, we propose a method to make inferences of the misclassification rate in the phenotype when gold standard samples are not available. Taking advantage of combining genetic data with EHR, we construct a maximum likelihood estimator to make inferences of the misclassification rate in cases/controls incorporating genotype information. Our approach is based on the insight that external GWAS without misclassification having similar ascertainment and design as our study can be a gold standard for the EHR GWAS. The effect sizes of genetic variants in our study should be similar to those in the external GWAS. Thus, we can estimate the misclassification rate by examining genotype information of dozens of disease-associated loci found by other large-scale GWAS.

Through extensive simulation studies and analysis of MGI data, we demonstrate that the proposed method can provide estimated misclassification rates with high accuracy under

scenarios of different misclassification rates, disease prevalence, and sample sizes. In addition, through evaluating different EHR-based case definition schemes, we can provide guidance about what is the best scheme to define cases for certain phenotypes. Moreover, knowing the misclassification rate in cases/controls will benefit the downstream analysis in many ways. First we can correct the effect size estimation with the misclassification rate using the formula derived by Neuhaus (1999) or Duffy's approach (2004) or using the iteratively reweighted least square algorithm (Magder and Hughes, 1997). In addition, the receiver operating characteristic (ROC) curve analysis can also be corrected using the misclassification rate (Zawistowski et al.,2017). Thus, combining our method with EHR-based GWAS and PheWAS analysis can help reduce the bias in the search across large numbers of phenotypes to broadly replicate and discover GWAS associations in EHR-based cohorts and enhance analysis of the genomic basis of human disease.

2.2 Method

2.2.1 Likelihood formulation

We have data of a binary trait and genotypes of n individuals. Let D_i denote true disease status and $\boldsymbol{G_i}$ denote the vector of genotypes of the ith individual. For each variant j, G_{ij} can either be genotype taking value {0, 1, 2} or imputed genotype dosage. Here we only consider known risk variants that are associated with disease. In other words, $\boldsymbol{G_i}$ only represents the genotype of known independent risk variants that have been discovered by previous large-scale GWAS. Let $\boldsymbol{X_i}$ denote a set of covariates (e.g., demographic variables and principal components for ancestry). We relate $\boldsymbol{D}$ to $\boldsymbol{G}$ and $\boldsymbol{X}$ through a generalized linear model with logistic link function,

$$logit(\pi_i) = \beta_0 + \sum_j \beta_j G_{ij} + \sum_k \gamma_k X_{ik} ,$$

where $\pi_i = Pr(D_i = 1|\boldsymbol{G_i}, \boldsymbol{X_i})$.

However, we do not observe the true response D directly but an error-corrupted response Y. Let α_0 and α_1 denote the specificity and sensitivity of measurement of Y respectively.Then $1-\alpha_0$ and $1-\alpha_1$ correspond to the probabilities of misclassification of the controls and cases in Y. Here we assume misclassification does not depend on the genotype and other covariates since the construction of the phenotype from EHR does not consider this information. In addition, we assume the misclassification is non-differentiable, which means α_0 and α_1 are the same for each sample. So that

$$1-\alpha_0 = Pr(Y_i = 1|D_i = 0, \boldsymbol{G_i}, \boldsymbol{X_i}) = Pr(Y_i = 1|D_i = 0)$$
$$1-\alpha_1 = Pr(Y_i = 0|D_i = 1, \boldsymbol{G_i}, \boldsymbol{X_i}) = Pr(Y_i = 0|D_i = 1).$$

Then the model for the observed phenotype Y can be calculated by taking the integral on the underlying true disease status D,

$$Pr(Y_i = 1|\boldsymbol{G_i}, \boldsymbol{X_i}) = \textstyle\sum_{d=0,1} Pr(Y_i = 1|D_i = d, \boldsymbol{G_i}, \boldsymbol{X_i}) * Pr(D_i = d|\boldsymbol{G_i}, \boldsymbol{X_i}).$$

Assuming the n samples are independent, the likelihood for observed data $\boldsymbol{Y}$, $\boldsymbol{G}$ and $\boldsymbol{X}$ should be

$$L(\boldsymbol{Y}, \boldsymbol{G}, \boldsymbol{X};\ \alpha_0, \alpha_1, \boldsymbol{\beta}, \boldsymbol{\gamma})$$

$$= \prod_i \{Pr(Y_i = 1|\boldsymbol{G_i}, \boldsymbol{X_i}) * Pr(\boldsymbol{G_i}, \boldsymbol{X_i})\}^{I(Y_i=1)} * \{Pr(Y_i = 0|\boldsymbol{G_i}, \boldsymbol{X_i}) * Pr(\boldsymbol{G_i}, \boldsymbol{X_i})\}^{I(Y_i=0)}$$

$$\propto \prod_i Pr(Y_i = 1|\boldsymbol{G_i}, \boldsymbol{X_i})^{I(Y_i=1)} * Pr(Y_i = 0|\boldsymbol{G_i}, \boldsymbol{X_i})^{I(Y_i=0)}.$$

Note that as $Pr(\boldsymbol{G_i}, \boldsymbol{X_i})$ does not contain the parameter α of interest, it can be factored out.

In the following analysis, for simplicity, we fix α_1 to be 1, assuming there is no misclassification of cases in the control group. We only focus on estimating α_0. One reason is that for most of the population-based diseases studies, the case/control ratio is not balanced. Usually, we have many

more controls than cases. As a consequence, even a small amount of misclassification in controls has a large impact on the association results while the misclassified cases typically have little impact. Therefore, we are more interested in the misclassified controls mixed in the cases.

Then the likelihood becomes

$L(\boldsymbol{Y}, \boldsymbol{G}, \boldsymbol{X};\ \alpha_0, \alpha_1, \boldsymbol{\beta}, \boldsymbol{\gamma})$

$$\propto \prod_i [Pr(D_i = 1|\boldsymbol{G_i}, \boldsymbol{X_i}) + (1 - \alpha_0)\Pr(D_i = 0|\boldsymbol{G_i}, \boldsymbol{X_i})]^{I(Y_i=1)} * [\alpha_0 Pr(D_i = 0|\boldsymbol{G_i}, \boldsymbol{X_i})]^{I(Y_i=0)}.$$

According to the logit model defined above, we have

$$(D_i = 1|\boldsymbol{G_i}, \boldsymbol{X_i}) = \frac{\exp(\beta_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik})}{1+\exp(\beta_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik})},\ Pr(D_i = 0|\boldsymbol{G_i}, \boldsymbol{X_i}) = \frac{1}{1+\exp(\beta_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik})}.$$

2.2.2 Incorporation of external GWAS information and parameter estimation

The model contains parameters, α_0, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$**.** Here, we assume the external GWAS with reliable phenotype to be the gold standard. Rather than estimating the parameter β, we borrow information of the effect sizes from external gold standard GWAS results and fix $(\beta_1, \beta_2, \dots, \beta_J)$ as the known effect sizes of the associated variants. By so doing, we are able to reduce the dimension of the parameters by the number of associated variants, J. This is valid under the assumption that effect sizes of disease-associated loci in our study are similar to those in the external gold standard GWAS.

The cohort of the external gold standard GWAS should be similar as our EHR study in terms of ancestry, gender, age, and other demographic factors. In order to get a reliable list of known variants, we suggest applying the following criteria: 1) filtering out ethnicity-specific variants; 2) selecting independent variants within each locus; 3) using effect sizes from replication studies

rather than discovery studies to avoid the ”winner's curse” (Bazerman and Samuelson, 1983; Lohmueller et al., 2003).

Then parameters can be estimated by maximizing the log-likelihood. This is a smooth nonlinear optimization problem. We solve the problem using "Augmented Lagrangian Adaptive Barrier Minimization Algorithm," which is implemented in an R package called "alabama." It has been shown that this algorithm can guarantee the convergence to the local minimum (Lange, 2004; Madsen et al., 2004). In order to increase the chance of achieving the global minimum, we randomly pick 10 initial points and choose the estimate that gives the maximum likelihood.

2.2.3 Inference of specificity

Once the estimation of the specificity, α_0 is obtained using the above method, we draw the inference regarding α_0 by testing for significance of $\widehat{\alpha_0}$. We test the point null hypothesis $\alpha_0 = 1$, which indicates there is no misclassification:

$$H_0: \alpha_0 = 1 \text{ vs. } H_1: \alpha_0 < 1.$$

The likelihood ratio test is chosen because it generally has better power than other tests like the Wald test and the score test (Casella and Berger, 2001). Furthermore, the distribution of the test statistic under the null hypothesis has been calibrated well when the test is done on the parameter space boundary (Self and Liang, 1987). The explicit formula of the log likelihood ratio test statistic is as follows:

$$\Lambda = 2 * \log\left(\frac{L(\widehat{\theta})}{L(\widehat{\theta_{H_0}})}\right),$$

where $\hat{\theta}$ is MLE of the parameters and $\widehat{\theta_0}$ is the MLE of the parameters under $\alpha_0 = 1$ constraint.

Since $\alpha_0 \in [0,1]$, we are testing a parameter on the boundary of the parameter space. Self and Liang (1987) theoretically showed that the above statistic Λ has an asymptotic mixed chi-square distribution under the null hypothesis,

$$\Lambda \sim 0.5{\chi_0}^2 + 0.5{\chi_1}^2.$$

Finally, the variance/covariance matrix of the parameter estimates can be approximated in a standard way using the inverse of the information matrix when the true α_0 is not 0 or 1. This matrix can be expressed in a relatively simple formula, which is provided in the Supplementary note.

2.3 Simulation studies

2.3.1 Estimation accuracy of parameters under scenarios with different sample sizes and disease prevalence

In the first section of the simulation, we examine the estimation accuracy of specificity in different settings. Genotypes of 300, 500, 1000 and 5000 samples are generated based on the allele frequencies (AFs) of 51 independent variants reported in a large-scale GWAS paper of AMD (Fritsche et al., 2016). We exclude one rare variant with extremely large effect size among the 52 reported variants (AF = 0.0001, OR = 20.3). A covariate, X, is generated from the normal distribution, $N(0,1)$. Then disease status is generated from a 1-0 Bernoulli distribution with probabilities $Pr(D_i = 1|\boldsymbol{G_i}, \boldsymbol{X_i}) = \frac{\exp(\beta_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik})}{1+\exp(\beta_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik})}$. Here, $(\beta_1, \beta_2, \dots, \beta_{51})$ are effect sizes referring to the same GWAS paper, and we set the effect size for covariate X, γ, to be 1. The intercept, β_0, which represents background disease prevalence, is set to be -2, -3 and -4. The exact marginal disease prevalence in the population based on AFs and background prevalence is not available. Because there are in total 3^{51} possible combinations of genotypes for 51 variants,

$$Pr(D_i = 1) = \sum_{G_i \in \kappa} Pr(D_i = 1|G_i) * Pr(G_i),$$

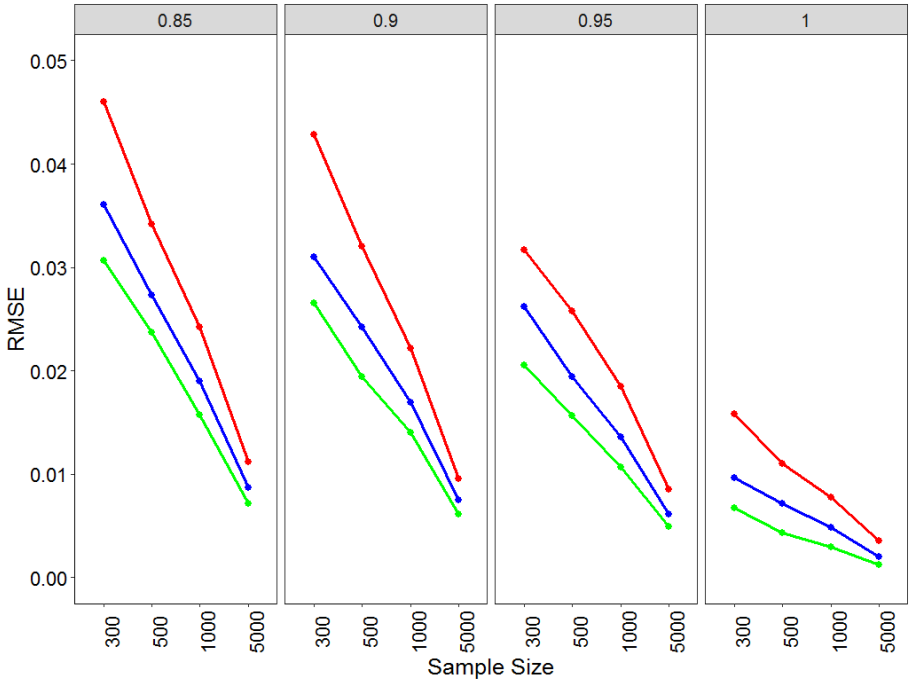
where κ is set of all combinations of genotypes. Therefore we calculate the empirical marginal disease prevalence by taking the mean of percentage of cases among 1000 simulated datasets. The marginal disease prevalence is ~4%, ~10% and ~15%, corresponding to background prevalence -4, -3 and -2 respectively.

Once we have data without misclassification, we contaminate the phenotype with misclassification rate $1 - \alpha_0$ in controls. More specifically, suppose there are n_0 true controls in the sample, we randomly draw n_0 phenotype from a $Bernoulli(1 - \alpha_0)$ distribution to replace our original phenotype of those true control samples. We set α_0 to be 0.85, 0.9, 0.95 and 1.

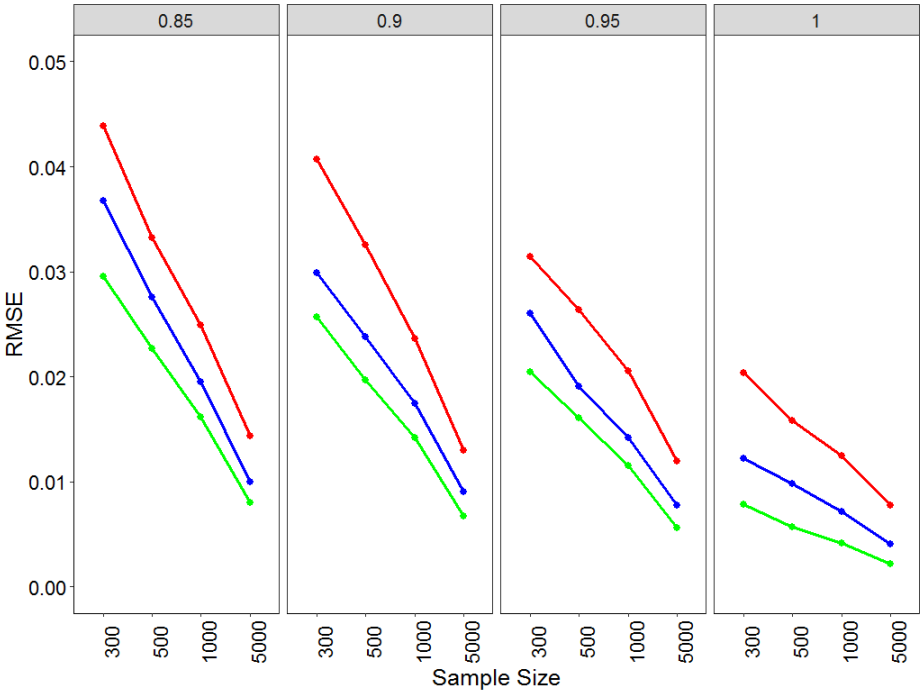
As stated in the previous part, the known effect sizes should be used as external information when estimating parameters. The ideal situation is that we know the true effect sizes of the variants. In reality, we estimate the effect size from GWAS. To mimic the real situation, we generate another independent dataset of 1000 cases and 4000 controls using the same settings as above but without contamination in the phenotype. Then effect sizes are estimated based on this dataset without misclassification. We apply both true effect sizes and estimated effect sizes to our model as the external gold standard information. Finally, we compare the results of using estimated effect sizes with the results of using true effect sizes to demonstrate the magnitude of bias that the noise in the estimated effect sizes introduces.

For each setting, we generate 1000 data sets and calculate the bias, variance and root-mean-square error (RMSE) to evaluate the estimation accuracy of α_0.

As can be seen from Figure 2.1, in general we get an accurate estimation of α_0 for different settings. In terms of RMSE, we have similar accuracy for different levels of specificity. As



(a)



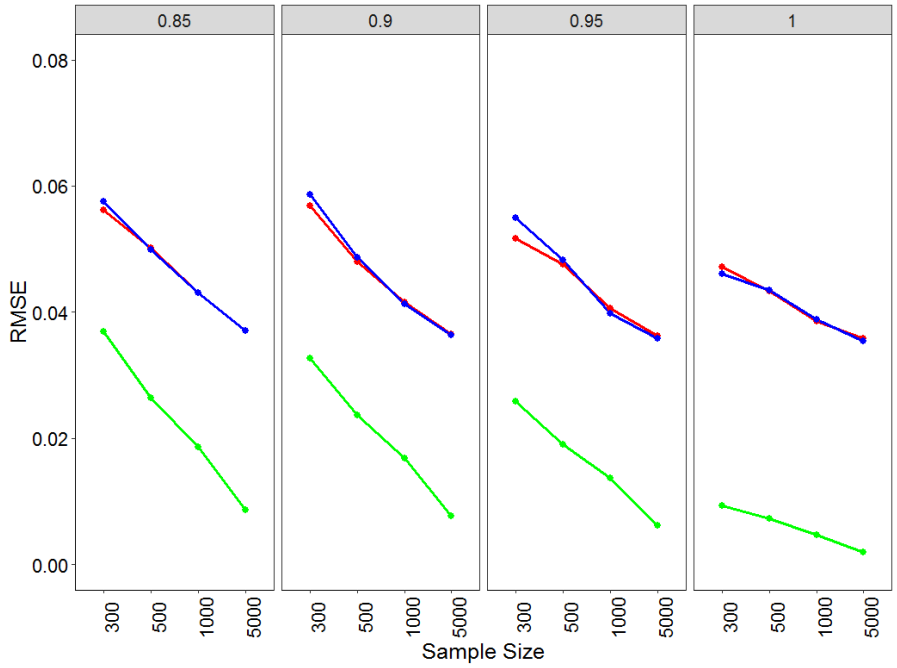
(b)

Figure 2.1(a,b): RMSE for estimation of specificity under different settings; (a) is using true effect sizes and (b) is using estimated effect sizes. The panel represents true specificity α_0. The red, blue and green line represent 15%, 10% and 4% disease prevalence respectively.

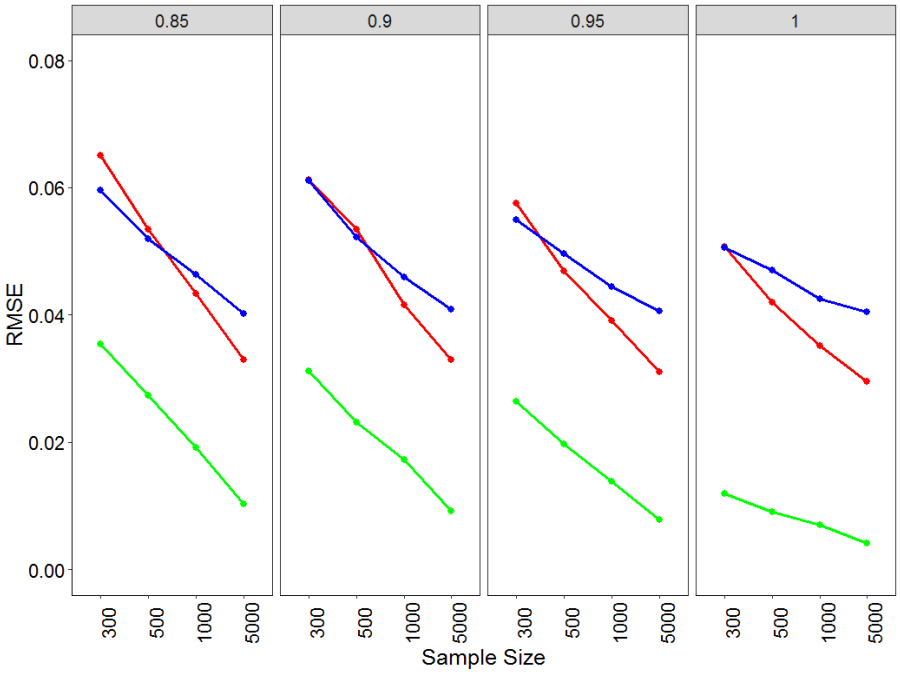
expected, increased sample size benefits the estimation accuracy. In addition, more accurate estimates occur in settings with lower disease prevalence. Moreover, in the comparison between applying true effect sizes and estimated effect sizes, we see higher RMSE in the latter settings. As Supplementary Table S2.1, the difference in bias explains a large proportion of the difference in RMSE. When applying estimated effect sizes in our likelihood, we introduce bias in the MLE of specificity and underestimate it.

In addition to assess the performance of our method, we compare our method with the method used in Tsoi et al. in 2017. Their method has a similar idea of using external GWAS as a gold standard, which estimates the misclassification rate by examining the median difference of risk allele frequencies (RAFs) between EHR data with external cohorts (Tsoi et al., 2017). The probability $Pr(D_i = 1|Y_i = 1)$ was calculated by $median(\frac{RAF_{case-EHR}-RAF_{ctrl-external}}{RAF_{case-external}-RAF_{ctrl-external}})$. However, they did not provide rigorous discussion and simulation for their method. We want to compare this method with our method through simulations. To make the comparison more comprehensive, for their method, we estimate $Pr(D_i = 1|Y_i = 1)$ not only by examining the median difference, but also by examining the mean difference. Then $\widehat{\alpha_0}$ is calculated based on Bayes' rule and compared with the estimation of our method.

As shown in Figure 2.2, Supplementary Table S2.2 and S2.3, our method outperforms the other two methods in all settings. For example, in the setting with 10% disease prevalence, the RMSE of our method is 1.5 to 18 times smaller than the other two methods under the same setting. As shown in Supplementary Table S2.2(b) and S2.3(b), the difference in RMSE between methods can be explained mainly by the discrepancies between the biases of the estimates. Both the mean and median method underestimate the parameter and the bias is greater than that of our method (Bias-our method = 0.0005 vs. Bias-mean = -0.035 vs. Bias-median = -0.035 for setting n=5000,



(a)



(b)

Figure 2.2(a,b): RMSE for estimation of specificity based on different methods; (a) is using true effect sizes and (b) is using estimated effect sizes. Results shown here are for settings with disease prevalence 10%. The panel represents true specificity α_0. The green line represents results using our method while red line and blue line represent method proposed by Tsoi et al. that examines mean and median differences respectively.

α_0 = 0.9, true effect sizes).

2.3.2 The effect of the number of associated variants examined on estimation accuracy

In real data analysis, it is not realistic to examine all variants that are associated with the disease. For example, the variant discovered by existing GWAS may not be genotyped or imputed in the EHR study. In addition, there is always some "missing heritability" that the existing GWAS do not have the power to uncover. In this section, we carry out the simulation to evaluate the performance of the proposed method when incorporating various amounts of information of the associated variants.

As stated above, fifty one variants are associated with the disease. Then we assume the external gold standard GWAS only provide association information of limited amount of those variants. Information of 1, 25 and 51 variants is examined in the model to estimate α_0.

In settings using 1 variant, in order to check if incorporating information of variants with larger effect size will benefit our estimation, we estimate α_0 using the variant with the largest effect size (effect size = 1.63) and compare the results of using a variant with moderate effect size (effect size = 0.17). In settings using 25 variants, the variants are randomly picked and fixed for every simulation. The remaining settings are kept the same as in Section 2.3.1.

Again, for each setting, we generate 1000 data sets. Then we use bias, variance and RMSE to evaluate the estimation accuracy of α_0.

Supplementary Table S2.4 shows the RMSE of estimated specificity when using different number of variants. In general, using information from all the 51 variants provides estimates with lowest RMSE compared to other settings. As the number of variants used decreases, we are

likely to have estimates with larger variance and bias which result in larger RMSE. Figure 2.3 shows the distribution of estimates when the sample size is 5000 with 10% disease prevalence and 0.9 specificity. Using one variant with the largest effect size (RMSE = 1.45E-04) has similar performance as using 25 variants (RMSE = 1.17E-04). Both outperform using one variant with a moderate effect size (RMSE = 4.84E-04). The RMSE for the latter one is over three times larger than using the one variant with the largest effect size. Later in Section 2.5, we will discuss how the number of variants examined impacts the estimation of specificity through a real data application.

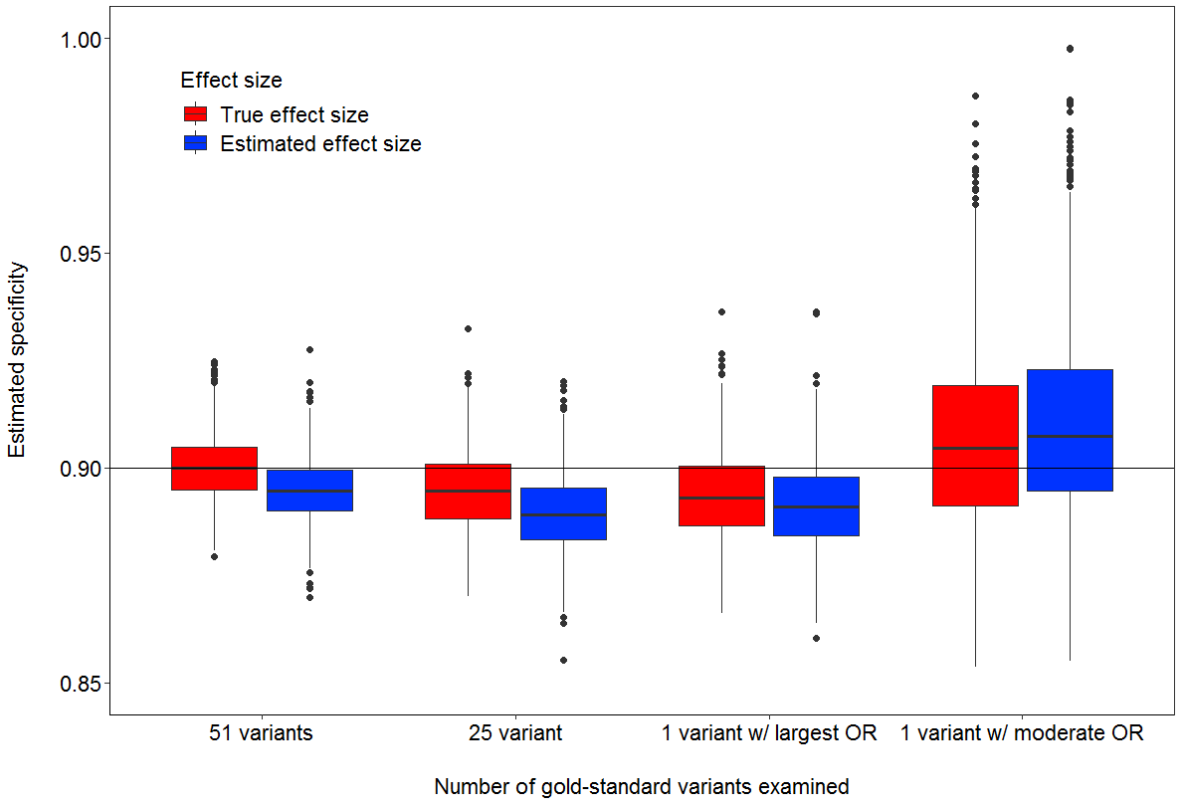


Figure 2.3: Distribution of estimated specificity by examining different number of variants. The table demonstrates the results under settings of 5000 samples, 0.9 specificity and 0.1 disease prevalence. The horizontal black line represents the true specificity, 0.9.

2.3.3 Distinguish "misclassified" samples due to the different liability thresholds from the truly misclassified ones

It is possible that due to the different liability threshold used to dichotomize cases and controls in the EHR GWAS study and other GWAS studies, a sample classified as a control in one study can be classified as a case in another study (Supplementary Figure S2.1). Those samples are not truly misclassified, but can bias the association result together with the truly misclassified ones. In this section, we mimic this situation to test if our method can distinguish these two set of samples.

According to the definition of the liability threshold model (Weissbrod et al., 2015), the liability of every individual i follows a normal distribution, $l_i \sim N(0,1)$. We generated the liability according to the model, $l_i = g_i + e_i$, where g_i and e_i are the genetic and environmental components of the liability, respectively. Cases are individuals with $l_i > t$, where t is the liability cutoff, i.e. the threshold, for a particular trait of interest defined by disease prevalence. If disease prevalence is K, t is given by $\Phi^{-1}(1-K)$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative probability density of the standard normal distribution.

In these simulations, each individual carries 51 causal variants with similar AF as the previous simulation and normally distributed effect sizes. Setting heritability to be 0.25, we generate environmental components from a normal distribution $N(0, {\sigma_e}^2)$, accordingly. Once the liability is obtained, case-control status is dichotomized by t satisfying disease prevalence K. Here we assume disease prevalence K_{EX} is 10% for the external gold standard GWAS from which we derived the estimated effect sizes. We assume the disease prevalence for the EHR study, K_{EHR}, to be 5%, 10%, 15% and 20%. The discrepancies between K_{EX} and K_{EHR} result in the first set of "misclassified" samples in the EHR. The second set of "misclassified" samples are generated by contaminating the phenotype using the same strategy as in the previous simulation with the

probability, 1-specificity. The specificity is set to be 0.85, 0.9, 0.95 and 1, and the sample size to 1000 and 5000. In all experiments, 1000 data sets are generated for each unique combination of settings. We examine whether our model can distinguish the second set of misclassified samples from the first set of by evaluating the estimation of α_0. Note that our method is constructed based on the logit model while data for this simulation is based on the liability threshold model. Thus, in addition to our main purpose, we can also demonstrate through this simulation whether our method is robust to model misspecification.

In Supplementary Table S2.5, comparing results for settings with no threshold difference, we conclude that if the true underlying model is a liability threshold model, we underestimate the specificity using our method that is constructed based on a logit model. Our method is sensitive to model misspecification.

If the liability threshold is lower in our EHR study than in the external GWAS where we borrow the information, the misclassification rate is overestimated and vice versa. This result implies that our model cannot distinguish the truly misclassified samples from those defined by a different dichotomizing liability threshold. The lower liability threshold in the EHR study results in more samples being classified as cases in EHR, our model treats those sets of samples as misclassified samples together with the truly misclassified ones, so that we overestimate the misclassification rate (Figure 2.4). When the external GWAS we use is assumed to have a different ascertainment than the EHR GWAS, the misclassification we estimate measures the total discrepancy in the phenotype between EHR and external GWAS. Thus, the interpretation of the estimated misclassification rate should change from "misclassification" to "misclassification and the difference of case ascertainments between EHR and external GWAS".

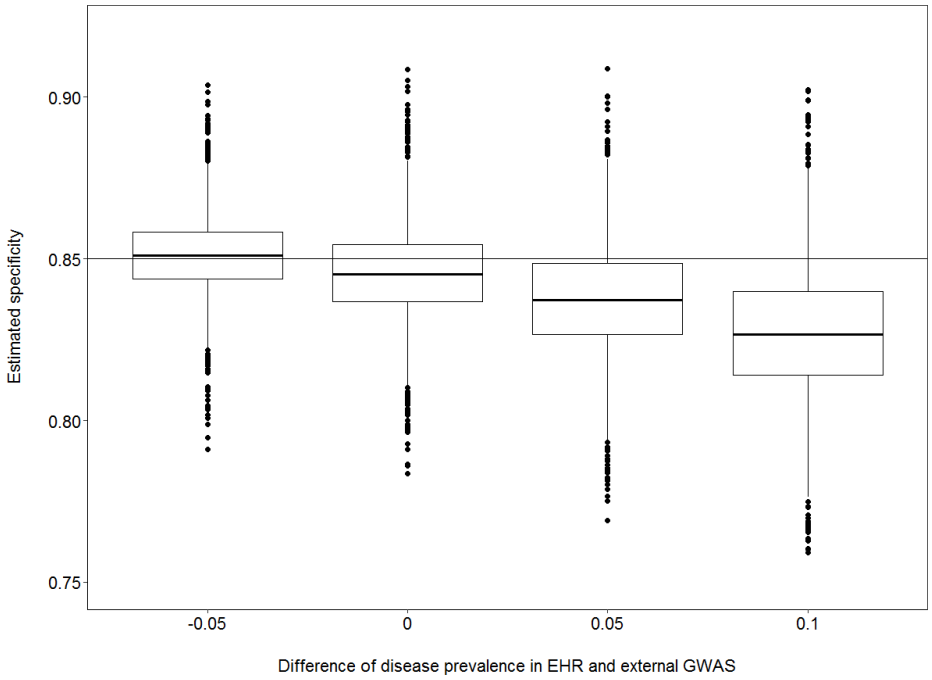


Figure 2.4: Distribution of estimated specificity when there is a liability threshold difference between EHR and external GWAS. The table shows the results under settings of 5000 samples and 0.85 specificity. The horizontal black line represents the true specificity, 0.85.

2.3.4 Type I error and power

In this section, we conduct extensive simulation studies to evaluate the performance of the likelihood ratio test. Data are generated the same way as in Section 2.3.1. Then we evaluate type I error in settings having α_0 equals 1 and power in other settings at significance level $\alpha = 0.01$ and 0.05.

The empirical type I error is shown in Table 2.1. At both 0.05 and 0.01 significance level, we can control the type I error well in all the 24 settings when applying the true effect sizes. However, we do suffer from the inconsistency of the estimator, when using estimated effect sizes. The type I error is inflated by 2 to 3 fold in those settings. To be more specific, we get more inflated results in settings with larger sample sizes, because the bias between true effect sizes and estimated effect sizes is amplified.

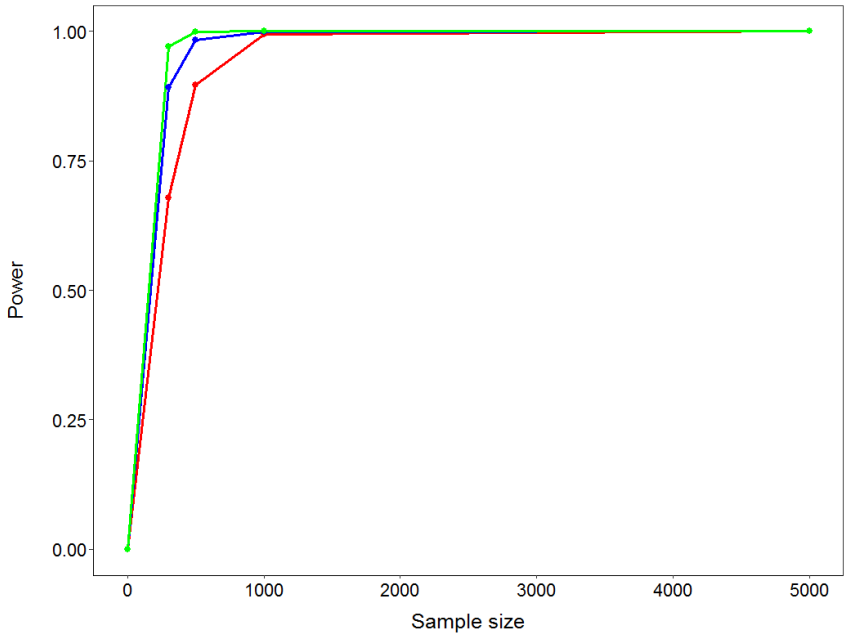
Table 2.1: Empirical type I error for likelihood ratio test testing null hypothesis. The null hypothesis refers to no misclassification in controls.

	Size / prevalence	alpha=0.05				alpha=0.01			
		300	500	1000	5000	300	500	1000	5000
True effect size	4%	0.05	0.05	0.04	0.05	0.01	0.01	0.01	0.01
	10%	0.05	0.05	0.05	0.05	0.02	0.01	0.01	0.01
	15%	0.05	0.05	0.05	0.05	0.01	0.01	0.01	0.01
Estimate effect size	4%	0.06	0.09	0.08	0.17	0.02	0.03	0.03	0.03
	10%	0.07	0.09	0.10	0.24	0.02	0.02	0.02	0.03
	15%	0.08	0.09	0.13	0.32	0.02	0.03	0.03	0.05

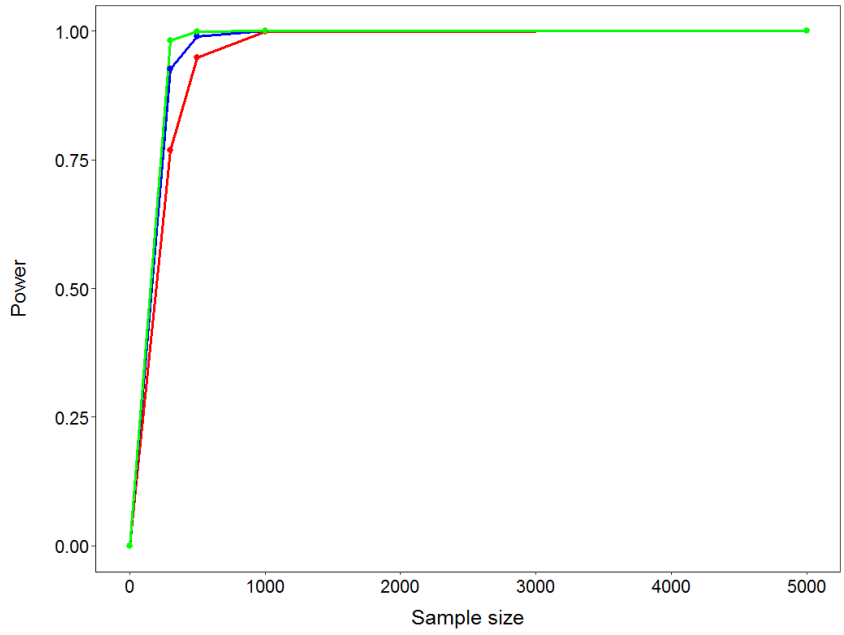
Here the external information of effect sizes is based on a cohort of only 5000 samples. In the real up-to-date GWAS, we could expect to get external estimated effect sizes based on a larger cohort. Thus we conduct a simulation using odds ratios calculated based on 20,000 samples. The type I error is shown in Supplementary Table S2.6. When estimated based on more samples, the external gold-standard effect sizes is closer to the truth. As a consequence, the type I error rate is better controlled. This simulation suggests that we should consider using effect sizes from larger studies that has better approximation to the true effect sizes generally to guarantee the control of type I error rate.

The power of the test is greater than 70% in most settings. Settings using the true or the estimated effect sizes perform similarly well, while settings with higher disease prevalence show a slight loss in power compared to settings with lower disease prevalence due to the less accurate estimation of specificity in the settings with high disease prevalence (Figure 2.5).

The results of power in the remaining set of simulations for all 48 simulation configurations are shown in Supplementary Tables S2.7.



(a)



(b)

Figure 2.5 (a,b): Power for likelihood ratio test under settings with specificity = 0.9 at 0.01 significance level; (a) is using true effect sizes and (b) is using estimated effect sizes. The red, blue and green line represent 15%, 10% and 4% disease prevalence respectively.

Based on results from Section 2.3.1 and 2.3.4, we conclude that noise in the estimated effect sizes does affect our estimation accuracy and the inference. That's why we emphasize the importance of study and risk variant selection in Section 2.2.

2.4 Application on MGI data

2.4.1 MGI data set

The Michigan Genomics Initiative is a collaborative study of the University of Michigan (UM) aiming to combine patient EHR with genetic information in order to gain novel biomedical insights.

Patients undergoing surgery at the UM Health System are invited to participate. Patients' biospecimens and their health information are collected during the surgical procedural period. Consenting patients are genotyped at 270K common variants on a customized Illumina Infinium HumanCoreExome-24 v1.0 array (Illumina, 2017). Then genotypes are imputed using the Haplotype Reference Consortium reference panel (Haplotype Reference Consortium, 2016). Meanwhile, phenotypes are constructed from ICD-9-CM billing based on a scheme implemented in the PheWAS R package. In total, there are 1,448 PheWAS codes with at least 20 cases for 18,267 unrelated European ancestry individuals that can be used for genome-wide association analysis in the Phase I MGI study. For more information on MGI, see https://precisionhealth.umich.edu/michigangenomics/.

The enriched information provides us a chance to accelerate the pace of genomic discovery. However, for some traits like type II diabetes, the attenuated effect sizes from previous MGI GWAS compared to effect sizes from other large-scale GWAS indicate potential misclassification in the MGI data (Supplementary Figure S2.2).

2.4.2 Estimation of misclassification rate in four phenotypes

To verify the existence of misclassification in MGI data and quantify it, we estimate the misclassification rate by fitting the model specified in the previous section for several traits. These traits showing evidence of misclassification in previous MGI GWAS include AMD, type II diabetes and psoriasis (Supplementary Figure S2.2(a-c)). We also apply our method to breast cancer which has not shown clear evidence of misclassification (Supplementary Figure S2.2(d)).

As described above, to fit the model, one critical step is to get gold standard information from external large GWAS results. Fifty-two independent variants have been previously reported to be associated with AMD by the International AMD Genomics Consortium (Fritsche et al., 2016). For our estimation, we use 44 variants that were available and passed filters (MAF>0.0001 and Rsq>0.3) in MGI data. For breast cancer, we incorporate results of five large-scale GWAS that reported 73 independent variants (Turnbull et al., 2010; Fletcher et al., 2011; Michailidou et al., 2013; Michailidou et al., 2015; Michailidou et al., 2017). For psoriasis, we combine results from five large-scale GWAS that reported 41 significant independent variants (Strange et al., 2010; Stuart et al., 2010; Tsoi et al., 2012; Tsoi et al., 2015; Yin et al., 2015). For T2D, we use 77 independent variants based on seven large-scale GWAS (Zeggini et al., 2008; Voight et al., 2010; Morris et al., 2012; Saxena et al., 2012; Consortium, Diabetes SAT2D, et al., 2014; Gaulton et al., 2015; Scott et al., 2017). Since some traits like AMD are highly correlated with age, we adjust for age in our model. The first four PCs are also adjusted in our model.

Once we estimate specificity, $\widehat{\alpha_0}$, we calculate the posterior misclassification rate, the misclassification rate in observed cases, based on the following Bayes' rule:

$$Pr(D_i = 1|Y_i = 1) = \frac{Pr(Y_i=1|D_i=1)*Pr(D_i=1)}{Pr(Y_i=1)}.$$

Here the marginal probability of observed cases $Pr(Y_i = 1)$ is estimated by the moment estimator. $Pr(D_i = 1)$ is estimated by the moment estimator derived based on $Pr(Y_i = 1) = Pr(D_i = 1) + (1 - \alpha_0) * (1 - Pr(D_i = 1))$.

The estimated misclassification rate in true controls is 0.3%, 1.5%, 0.8% and 4.8% for AMD, breast cancer, psoriasis and T2D resulting in 39.5%, 9.0%, 56.5% and 37.8% misclassified samples, respectively, in the observed cases (Table 2.2). The Bonferroni-corrected significance threshold is derived as p-value < 0.05/4 = 0.0125 to account for examination of the four traits. The results for AMD, T2D and psoriasis are significant ($p\text{-value}_{AMD}$ = 2.10E-10; $p\text{-value}_{T2D}$ = 4.70E-15; $p\text{-value}_{psoriasis}$ = 5.63E-15) while the result for breast cancer is not significant ($p\text{-value}_{breast\ cancer}$ = 6.71E-2). This results shows concordance with the effect sizes comparison with the external GWAS (Supplementary Figure S2.2(a-d)). We also estimate the specificity by examining half of the most significant variants as well as one variant with largest effect size (Table 2.4). The impact of variants selection on the estimation and the source of misclassification of these diseases will be discussed in Section 2.5.

Table 2.2: Estimated misclassification rate in 4 traits.

	Cases	Controls	Estimated (1-α_0)	P-value	Misclassification rate in observed cases
AMD	119	16516	0.3%	2.10E-10	39.5%
Breast cancer	1136	6884	1.5%	6.71E-2	9.0%
Psoriasis	231	15612	0.8%	5.63E-15	56.5%
T2D	1974	14848	4.8%	4.70E-15	37.8%

2.4.3 Comparison of different case definition schemes for AMD

Patients' clinical information is gathered during every hospital visit. Therefore, we usually have more than one encounter of each phenotype for the sample. On average, samples had about 20 to 30 encounters for one trait.

The longitudinal feature of the data raises the question of the best way to define a case. In the previous MGI GWAS, we defined a case if one had more than 2 encounters. However, this may not be the best case definition scheme. Usually, patients with a larger number of encounters are more likely to be a true case. So in this section, taking AMD as an example, we try to estimate the misclassification rate for different case-definition schemes and to find the best encounter cutoff to define a case.

Table 2.3 and Supplementary Table S2.8 show the estimated misclassification rate in observed AMD cases when using 1 to 13 as encounter cutoff to define a case. Here, to guarantee the reliability of the estimation, we only examine the encounter cutoffs that define more than 30 cases. As the cutoff becomes more stringent, we get cohorts of fewer cases with a lower misclassification rate. Based on the Holm's sequential Bonferroni method (1979), we define the significance threshold for the sequential tests. The misclassification in the phenotype is not significant with encounter cutoff greater than 7 (Supplementary Table S2.8). Multiplying observed cases with the 1-misclassification rate, we have the expected number of correctly classified cases, which shows how much we gain when using more stringent cutoffs. For example, if releasing the criterion from 3 to 2 we get 33 new cases, of which only 13 are correctly specified. If we release the criterion from 4 to 3, we get 12 more cases that half of them are correctly specified. As the encounter cutoff becomes more stringent, the proportion of correctly classified samples in the newly defined samples increases.

Using the phenotype with less misclassification for GWAS, we expect to get the estimation of the effect size with smaller bias. Here we examine the top two common variants with the largest effect sizes, rs3750846 around gene *ARMS2* and rs10922109 around gene *CFH*, to check if their results are converge to the external GWAS when we use refined phenotype (Fritsche et al., 2016).

In terms of the minor allele frequency (MAF) in the cases and the estimated effect size, MGI results converge to the external gold-standard as we increase encounter cutoffs (Supplementary Table S2.9), which is concordant with the decreasing misclassification rate in the phenotype.

The results imply that as we use a higher cutoff of encounters to define a case, we have a more reliable set of cases. A caveat is that the number of cases also decreases as we refine the phenotype, which results in the unbalanced case-control ratios invalidating asymptotic assumptions of the logistic regression. Thus using phenotype with small case count, even though it is refined, we are not able to achieve the convergence of the effect sizes to the external AMD GWAS for the variants having low AFs (Supplementary Figure S2.3). In Section S 2.5, we will discuss several possible solutions of how to increase the power of GWAS with refined phenotype that has small number of cases and how to correct the effect size estimation directly with the misclassification rate that our method provides.

Table 2.3: Estimated misclassification rate in observed cases when using different encounter cutoffs to define an AMD case.

Cutoff	Cases	Controls	Estimated ($1-\alpha_0$)	Misclassification rate in observed cases	Average number of correctly classified samples
≥1 encounter	144	16516	0.40%	47.40%	76
≥2 encounters	119	16516	0.30%	39.50%	71
≥3 encounters	86	16516	0.16%	32.30%	58
≥4 encounters	74	16516	0.12%	28.50%	53
≥5 encounters	65	16516	0.10%	26.62%	48
≥6 encounters	55	16516	0.10%	29.49%	39
≥7 encounters	49	16516	0.06%	20.26%	39
≥8 encounters	45	16516	0.04%	14.92%	38
≥9 encounters	44	16516	0.04%	15.50%	37
≥10 encounters	42	16516	0.04%	16.40%	35
≥11 encounters	38	16516	0.02%	9.50%	34
≥12 encounters	37	16516	0.02%	10.60%	33
≥13 encounters	33	16516	0.02%	8.57%	30

2.5 Discussion

In this chapter, we have developed a method for estimating the misclassification rate of disease status using genotype information. The method can provide estimates with high accuracy and has the advantage of simplicity that avoids identifying gold standard samples.

In the comparison with the method proposed by Tsoi et al (2017), our method demonstrates higher accuracy in the estimation of specificity. The outperformance of our method is more apparent in practice. Our method requires external information of odds ratios or effect sizes while Tsoi's method requires RAFs in cases/controls. However, the background disease prevalence is unknown so the exact RAFS cannot be calculated. Instead, they are approximated based on odds ratios and population AFs. As our simulation shows, using approximated RAFs lead to an even less accurate estimation than using exact RAFs (Supplementary Table S2.4). In contrast, odds ratios are generally reported, and thus our method has an obvious advantage regarding estimation accuracy when exact RAFs are not available in the real data analysis.

Our method is easy to apply and useful, although potential failures of the assumptions involved should be borne in mind. First, our model is based on the assumption that the effect sizes in the EHR GWAS are very similar to those in the external GWAS. As demonstrated in the simulation, we get a good estimate when the gold-standard effect sizes we borrow are not far from the true effect sizes. Thus, it is necessary to carefully consider the plausibility of this assumption in the context of the individual study, especially for a phenotype that is highly correlated with demographic factors. For example, AMD is highly correlated with age, and the distribution of age in the external GWAS cohorts may be different from the EHR cohort, which may lead to different disease-genetic associations. Even though we allow the control for demographic covariates in our model, the non-linear relationship between the disease onset and age may still

bias the estimation of the specificity when we ignore the cohorts' demographic discrepancy. To avoid this problem, we should always select external GWAS that have similar cohorts as the EHR cohort to be the gold standard. Second, we should be careful when selecting the associated variants. In our analysis of MGI data, we only borrowed information from reliable large-scale GWAS with more than 10k samples. One may consider using a large database like the GWAS catalog from which to draw information. However, in this database, results are not well harmonized in that GWAS of different phenotypes may be combined into one trait. Therefore we suggest using them with a cautious screening of the variants.

Within these variant inclusion restrictions, the simulation result in Section 2.3.2 suggests that we should consider using as many associated variants that pass the filter as possible; and variants with larger effect sizes should have higher priority to be included. To examine the consistency of the estimated specificity with different number of variants incorporated into the model, we can conduct a sensitivity analysis in MGI. When one of the associated variants has an effect size that is significantly larger than the others, like AMD, the estimation is consistent, since the effect of the other variants may be masked by that variant. When all the associated variants have moderate effect sizes, the results are concordant with the conclusion of the previous simulation analysis that the specificity will be slightly overestimated when the information of some of the reliable associated variants is not examined (Table 2.4). Thus, our method demonstrates the utility for diseases whose association with genetic variation has been investigated through large-scale GWAS. However, our method is not applicable for diseases that do not have well-established GWAS. For those diseases, we may consider first selecting potential associated variants based on EHR GWAS, and then using the EM algorithm proposed by Magder and Hughes (1997) to estimate the misclassification rate without incorporating external effect sizes.

Table 2.4: Estimated specificity in 4 traits by examining different number of variants. Specificity is estimated using all associated variants, half of the variants with the largest effect sizes and the variant with the largest effect size.

	All variants	Half variants	One variant
AMD	0.3%	0.3%	0.3%
Breast cancer	1.5%	1.7%	2.0%
Psoriasis	0.8%	0.9%	1.0%
T2D	4.8%	4.9%	4.9%

In addition to carefully selecting the variant list, the uncertainty in the estimation of external gold-standard effect sizes should be considered. Usually, the effect sizes are reported in multiple GWAS that have different sample sizes. We take the average of the effect sizes in multiple external GWAS and fix it in our model. In future work, to better address the uncertainty in the estimation of effect sizes, we can take the weighted average of effect sizes, by weighting each GWAS result based on their sample sizes. In addition, instead of fixing external effect sizes in our model, we can estimate these effect sizes together with the specificity and constrain the estimation within the confidence intervals in the external gold-standard GWAS.

Moreover, once we estimate the misclassification rate by examining a reliable set of variants, it can be used as a quality metric for refining the phenotype. Typically, the effect size estimation in EHR GWAS with the refined phenotype is expected to converge to that in the external gold-standard GWAS. However, the reduced number of cases and the unbalanced case/control ratios of the refined phenotype cause a wider range of uncertainty in the effect size estimation. As a consequence, we may not necessarily see the convergence of the refined EHR GWAS to the external GWAS, especially for variants with low AF. In order to retain the power of uncovering disease-genetic associations, instead of using the traditional logistic regression, we can conduct association tests that account for unbalanced case/control ratio in the phenotype (Zhou et al., 2018). In addition, rather than refining the phenotype, we may consider correcting estimated

effect sizes directly with the estimated misclassification rate using the formulas derived by Neuhaus (1999) or Duffy (2004), or using the iteratively reweighted least square algorithm (Magder and Hughes, 1997). Figure 2.6 shows the estimated effect sizes of the two top associated variants, rs3750846 and rs10922109, using different methods. Either using refined phenotype (≥ 7 encounters) or correcting effect sizes directly with Duffy's method provides us with results that are closer to the external GWAS than those estimated from a misclassified cohort (≥ 2 encounters).

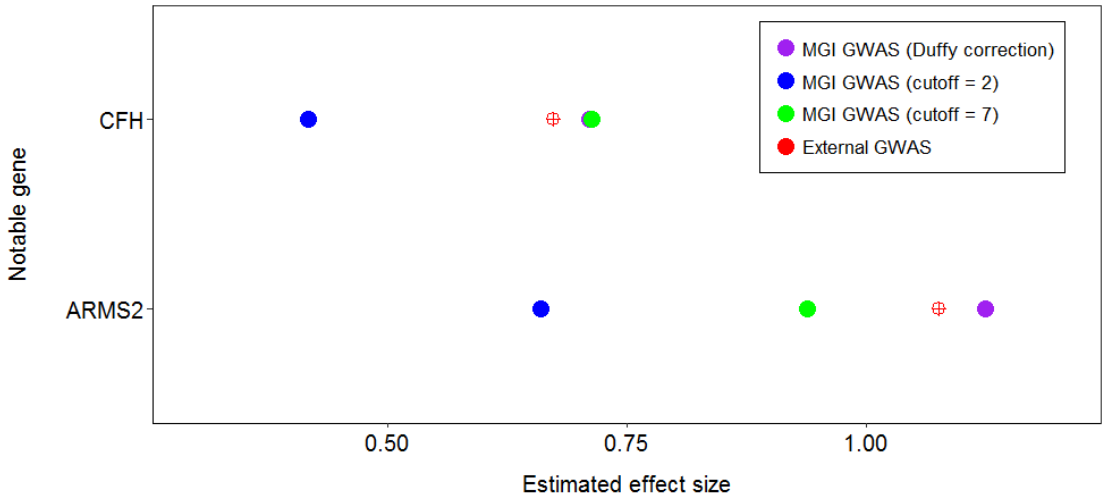


Figure 2.6 Estimated effect sizes in external GWAS and EHR GWAS with refined phenotype or with Duffy's correction. Here two top associated variants, rs3750846 (ARMS2) and rs10922109 (CFH), are examined (Fritsche et al., 2016). The red targets represent their effect sizes in external GWAS. The blue dots represent estimation using misclassified samples that are defined by encounter cutoff = 2. The green dots represent estimation using misclassified samples that are defined by encounter cutoff = 7, which have no significant misclassification. The purple dots represent effect sizes corrected from the blue dots using Duffy's method.

In the application of four traits in MGI data, significant misclassification was detected in AMD, psoriasis and T2D while no significant misclassification was detected in breast cancer. The misclassification can occur either during the translation from ICD codes to the dichotomized phenotypes or during the assignment of ICD codes to patients. First, the ICD codes are typically

not surrogates for binary phenotypes; in other words, multiple ICD codes may describe the same diagnosis. Thus, the complexity in the translation from ICD codes to phenotypes may lead to the potential misclassification. For example, for T2D, we combine 10 different ICD codes having "type II diabetes mellitus" in their description into one phenotype. However, some of these ICD codes may refer to types of T2D that should be analyzed separately, like "250.10 - Diabetes with ketoacidosis" and "250.20 - Diabetes with hyperosmolarity." In addition, the descriptions of some ICD codes are ambiguous. For example, the ICD code 696 for psoriasis has the description, "psoriasis and similar disorders," which may classify patients with diseases other than psoriasis as a case. For AMD, the ICD codes do not indicate the stage of the disease, so that patients at early or moderate stage of AMD may be classified as a case in EHR while the external GWAS are typically conducted on advanced AMD. Moreover, as discussed by O'Malley et al. (2005), errors in ICD code can occur in every step of them being assigned to patients, including the communication between patients and clinicians, the clinicians' knowledge of the disease as well as intentional code errors, like upcoding. Upcoding, here, means that codes of higher reimbursement value may be assigned due to some reimbursement purpose. It misrepresents the true condition of the patient, which may cause controls being misclassified as cases. O'Malley et al. also assert that a disease for which tests have high sensitivity will have higher diagnostic accuracy and smaller error in the ICD codes. This is why breast cancer typically has higher diagnostic accuracy and does not show significant misclassification in the MGI data.

Ultimately, by using our method we are able to detect misclassification between cases and controls and correct biased association analysis in EHR GWAS. However, there are still other types of misclassification that limit the usage of EHR. For example, despite the fact that T1D and T2D arise from different etiologies, it is hard to distinguish these two diseases based on ICD-9-

CM codes (Kho et al., 2011; Richesson et al., 2013). Since both diseases are characterized by high blood glucose level and share similar treatment, only considering additional information of diagnostic lab tests and treatment is not adequate to distinguish them. It should be more useful to incorporate genotype information given that these two diseases have different associated variants. Future development of methods dealing with misclassification between cases like T1D and T2D using genotype information will bring about more powerful genetic discovery research using EHR data.

In conclusion, we have proposed a method that enables the evaluation of new EHR-based case definition schemes and the correction of estimates of disease effect sizes and other association measures when phenotypes are misclassified. This method has limitations that need further investigation, including the inefficient process of risk variants selection and the disregarding of the misclassification between different phenotypes. Nevertheless, the method can reduce the bias in the search of disease-genetic associations and enhances the power of uncovering novel and reliable genomic basis of human diseases.

Supplements

Supplementary note: Estimating the variance of the parameter estimates when true specificity is not on the boundary.

When the model is correctly specified and the true parameter is not on the boundary, the variance/covariance matrix of the parameter estimates can be approximated by the inverse of the observed information matrix. This matrix is the negative of the second derivative of the log likelihood function evaluated at the maximum likelihood estimates.

Recall the log-likelihood has this form ,

$$l(\boldsymbol{Y}, \boldsymbol{G}, \boldsymbol{X};\ \alpha_0, \alpha_1, \boldsymbol{\beta}, \boldsymbol{\gamma})$$

$$\propto \sum_i Y_i \log\left(\frac{e^{\beta_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik}}}{1+e^{\beta_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik}}} + \frac{1-\alpha_0}{1+e^{\beta_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik}}}\right)$$

$$+ (1-Y_i)\log\left(\frac{\alpha_0}{1+e^{\beta_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik}}}\right)$$

Here, we assume $(\beta_1, \dots, \beta_J)$ are known parameters indicating effect sizes of the kwon diseased associated variants. Then the parameters we estimate are $\theta = (\alpha_0, \beta_0, \gamma_1, \dots, \gamma_K)$. To simplify the notation, we use $\gamma_0$to represent β_0 and $X_{i0} = (1, \dots, 1)$. So that θ becomes $(\alpha_0, \gamma_0, \gamma_1, \dots, \gamma_K)$.

Let $n_0 = \sum_i (1 - Y_i)$, the second derivatives of the likelihood function regarding the parameters are:

$$\frac{\partial l}{\partial \alpha_0^2} = -\frac{n_0}{\alpha_0^2} - \sum_i \frac{Y_i}{\left(e^{\gamma_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik}} + 1 - \alpha_0\right)^2},$$

$$\frac{\partial l}{\partial \alpha_0 \partial \gamma_k} = \sum_i \frac{Y_i X_{ij} e^{\gamma_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik}}}{(e^{\gamma_0+\sum_j \beta_j G_{ij}+\sum_k \gamma_k X_{ik}} + 1 - \alpha_0)^2},$$

$$\frac{\partial l}{\partial \gamma_m \partial \gamma_n} = \sum_i \frac{(1-\alpha_0) Y_i X_{im} X_{in} e^{\gamma_0 + \sum_j \beta_j G_{ij} + \sum_k \gamma_k X_{ik}}}{(e^{\gamma_0 + \sum_j \beta_j G_{ij} + \sum_k \gamma_k X_{ik}} + 1 - \alpha_0)^2} - \sum_i \frac{X_{im} X_{in} e^{\gamma_0 + \sum_j \beta_j G_{ij} + \sum_k \gamma_k X_{ik}}}{\left(e^{\gamma_0 + \sum_j \beta_j G_{ij} + \sum_k \gamma_k X_{ik}} + 1\right)^2}.$$

Let $\hat{\theta} = (\widehat{\alpha_0}, \widehat{\gamma_0}, \widehat{\gamma_1}, \ldots, \widehat{\gamma_K})$ be the MLE. Then the observed information matrix is:

$$J_n = -\frac{\partial l^2}{\partial \theta \partial \theta'}\Big|_{\theta=\hat{\theta}} = \left[\begin{array}{c|c} A & B^T \\ \hline B & C \end{array}\right];$$

$$A = \frac{n_0}{\widehat{\alpha_0^2}} + \sum_i \frac{Y_i}{\left(e^{\widehat{\gamma_0} + \sum_j \beta_j G_{ij} + \sum_k \widehat{\gamma_k} X_{ik}} + 1 - \widehat{\alpha_0}\right)^2};$$

$B = -X^T H$, where $H_i = \frac{Y_i e^{\widehat{\gamma_0} + \sum_j \beta_j G_{ij} + \sum_k \widehat{\gamma_k} X_{ik}}}{(e^{\widehat{\gamma_0} + \sum_j \beta_j G_{ij} + \sum_k \widehat{\gamma_k} X_{ik}} + 1 - \widehat{\alpha_0})^2}$;

$C = X^T V_1 X - X^T V_2 X$, where V_1 is a diagonal matrix with diagonal element

$v_{1i} = \frac{e^{\widehat{\gamma_0} + \sum_j \beta_j G_{ij} + \sum_k \widehat{\gamma_k} X_{ik}}}{\left(e^{\widehat{\gamma_0} + \sum_j \beta_j G_{ij} + \sum_k \widehat{\gamma_k} X_{ik}} + 1\right)^2}$ and V_2 is a diagonal matrix with diagonal element $v_{2i} =$

$\frac{(1-\widehat{\alpha_0}) Y_i e^{\widehat{\gamma_0} + \sum_j \beta_j G_{ij} + \sum_k \widehat{\gamma_k} X_{ik}}}{(e^{\widehat{\gamma_0} + \sum_j \beta_j G_{ij} + \sum_k \widehat{\gamma_k} X_{ik}} + 1 - \widehat{\alpha_0})^2}$.

Thus when the true specificity α_0 is not 0 or 1 and the effect sizes of known disease associated variants are correctly specified, the variance/covariance matrix of the estimates can be approximated by $var(\theta) = {J_n}^{-1}$.

Supplementary Table S2.1 (a,b): (a) MSE and (b) bias for estimation of specificity under different settings.

(a)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	9.42E-04	7.05E-04	4.22E-04	4.56E-05	1.30E-03	9.62E-04	6.84E-04	9.24E-05	2.12E-03	1.83E-03	1.00E-03	2.51E-04
	500	5.61E-04	3.77E-04	2.45E-04	1.86E-05	7.45E-04	5.85E-04	3.78E-04	5.08E-05	1.17E-03	1.03E-03	6.65E-04	1.21E-04
	1000	2.47E-04	1.95E-04	1.14E-04	8.52E-06	3.62E-04	2.87E-04	1.86E-04	2.36E-05	5.89E-04	4.90E-04	3.40E-04	5.95E-05
	5000	5.09E-05	3.71E-05	2.46E-05	1.50E-06	7.49E-05	5.61E-05	3.78E-05	3.92E-06	1.25E-04	9.11E-05	7.21E-05	1.24E-05
Estimate effect size	**300**	8.74E-04	6.58E-04	4.18E-04	6.17E-05	1.35E-03	8.91E-04	6.76E-04	1.49E-04	1.93E-03	1.65E-03	9.87E-04	4.15E-04
	500	5.15E-04	3.88E-04	2.58E-04	3.24E-05	7.61E-04	5.67E-04	3.63E-04	9.59E-05	1.11E-03	1.06E-03	6.94E-04	2.50E-04
	1000	2.62E-04	2.00E-04	1.33E-04	1.74E-05	3.81E-04	3.06E-04	2.01E-04	5.10E-05	6.19E-04	5.57E-04	4.22E-04	1.55E-04
	5000	6.38E-05	4.57E-05	3.16E-05	4.66E-06	9.95E-05	8.24E-05	6.06E-05	1.68E-05	2.05E-04	1.70E-04	1.43E-04	6.07E-05

(b)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	4.2E-03	2.4E-03	1.5E-03	-3.1E-03	2.6E-03	2.1E-03	8.7E-04	-4.6E-03	2.8E-03	1.7E-03	3.8E-04	-7.7E-03
	500	1.7E-03	6.9E-04	1.8E-03	-2.0E-03	2.6E-03	1.9E-03	4.2E-04	-3.6E-03	1.6E-03	9.2E-04	6.5E-04	-5.5E-03
	1000	7.6E-04	7.6E-04	5.9E-04	-1.4E-03	1.4E-03	3.5E-04	1.5E-03	-2.3E-03	-3.1E-04	1.0E-03	1.3E-03	-4.0E-03
	5000	2.4E-04	1.2E-04	-1.7E-04	-6.3E-04	-9.3E-07	-2.7E-05	-9.2E-05	-1.1E-03	2.0E-04	3.1E-04	1.5E-04	-1.9E-03
Estimate effect size	**300**	-1.6E-03	-6.1E-04	-1.3E-03	-3.9E-03	5.5E-04	-5.3E-03	-2.7E-03	-6.6E-03	-6.6E-03	-7.6E-03	-6.4E-03	-1.2E-02
	500	-1.1E-03	-1.0E-03	-1.1E-03	-3.0E-03	-3.7E-03	-3.9E-03	-3.5E-03	-5.4E-03	-6.8E-03	-8.5E-03	-7.1E-03	-9.4E-03
	1000	-2.3E-03	-2.3E-03	-2.1E-03	-2.3E-03	-4.1E-03	-4.9E-03	-4.3E-03	-4.4E-03	-7.9E-03	-9.3E-03	-8.5E-03	-8.2E-03
	5000	-2.8E-03	-2.8E-03	-2.6E-03	-1.5E-03	-5.5E-03	-5.1E-03	-4.9E-03	-3.1E-03	-9.5E-03	-9.1E-03	-8.7E-03	-6.2E-03

Supplementary Table S2.2 (a,b): (a) MSE and (b) bias for estimation of specificity from the method proposed by Tsoi et al. by examining the mean RAFs difference.

(a)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	1.79E-03	1.64E-03	1.20E-03	6.49E-04	3.15E-03	3.23E-03	2.67E-03	2.22E-03	8.47E-03	8.76E-03	8.78E-03	9.03E-03
	500	1.24E-03	1.03E-03	7.99E-04	4.55E-04	2.51E-03	2.30E-03	2.26E-03	1.88E-03	7.25E-03	8.01E-03	7.93E-03	8.03E-03
	1000	7.73E-04	5.88E-04	4.40E-04	2.86E-04	1.86E-03	1.72E-03	1.64E-03	1.48E-03	6.54E-03	6.79E-03	7.42E-03	7.61E-03
	5000	3.00E-04	2.68E-04	2.06E-04	1.69E-04	1.37E-03	1.33E-03	1.32E-03	1.28E-03	6.01E-03	6.29E-03	6.65E-03	7.06E-03
Estimate effect size	**300**	2.68E-03	2.37E-03	1.69E-03	1.03E-03	4.23E-03	3.75E-03	3.31E-03	2.57E-03	8.57E-03	9.48E-03	9.02E-03	9.06E-03
	500	1.59E-03	1.42E-03	1.08E-03	5.96E-04	2.85E-03	2.86E-03	2.20E-03	1.76E-03	6.39E-03	6.90E-03	7.78E-03	6.82E-03
	1000	9.38E-04	7.49E-04	5.81E-04	3.09E-04	1.88E-03	1.72E-03	1.53E-03	1.24E-03	5.66E-03	5.63E-03	6.20E-03	6.05E-03
	5000	2.63E-04	2.28E-04	1.73E-04	1.09E-04	1.08E-03	1.08E-03	9.64E-04	8.74E-04	4.73E-03	5.09E-03	5.15E-03	5.29E-03

(b)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	-1.4E-02	-1.2E-02	-1.0E-02	-1.1E-02	-3.1E-02	-3.4E-02	-3.2E-02	-3.2E-02	-7.2E-02	-7.6E-02	-7.8E-02	-8.1E-02
	500	-1.4E-02	-1.2E-02	-1.2E-02	-1.1E-02	-3.4E-02	-3.4E-02	-3.5E-02	-3.3E-02	-7.4E-02	-7.9E-02	-8.0E-02	-8.1E-02
	1000	-1.3E-02	-1.1E-02	-1.2E-02	-1.1E-02	-3.5E-02	-3.4E-02	-3.4E-02	-3.4E-02	-7.5E-02	-7.7E-02	-8.2E-02	-8.3E-02
	5000	-1.4E-02	-1.3E-02	-1.2E-02	-1.2E-02	-3.5E-02	-3.5E-02	-3.5E-02	-3.5E-02	-7.6E-02	-7.8E-02	-8.1E-02	-8.3E-02
Estimate effect size	**300**	-8.1E-03	-9.0E-03	-6.6E-03	-6.8E-03	-2.9E-02	-2.8E-02	-2.4E-02	-2.4E-02	-6.0E-02	-6.7E-02	-6.6E-02	-6.9E-02
	500	-9.4E-03	-8.6E-03	-6.4E-03	-6.1E-03	-2.7E-02	-3.0E-02	-2.7E-02	-2.6E-02	-6.0E-02	-6.5E-02	-7.1E-02	-6.5E-02
	1000	-8.6E-03	-9.6E-03	-7.8E-03	-6.6E-03	-3.0E-02	-2.9E-02	-2.7E-02	-2.5E-02	-6.5E-02	-6.5E-02	-6.9E-02	-6.9E-02
	5000	-1.0E-02	-9.5E-03	-8.7E-03	-7.5E-03	-2.9E-02	-3.0E-02	-2.8E-02	-2.8E-02	-6.7E-02	-6.9E-02	-7.0E-02	-7.1E-02

Supplementary Table S2.3 (a,b): (a) MSE and (b) bias for estimation of specificity from the method proposed by Tsoi et al. by examining the median RAFs difference.

(a)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	1.87E-03	1.67E-03	1.24E-03	6.07E-04	3.31E-03	3.44E-03	3.01E-03	2.12E-03	8.85E-03	9.00E-03	9.33E-03	9.57E-03
	500	1.16E-03	9.86E-04	8.52E-04	4.13E-04	2.49E-03	2.37E-03	2.32E-03	1.89E-03	7.28E-03	8.16E-03	7.60E-03	8.13E-03
	1000	7.18E-04	5.72E-04	4.56E-04	2.85E-04	1.86E-03	1.71E-03	1.58E-03	1.51E-03	6.54E-03	6.92E-03	7.14E-03	7.25E-03
	5000	2.93E-04	2.68E-04	2.05E-04	1.60E-04	1.38E-03	1.32E-03	1.28E-03	1.26E-03	5.88E-03	6.10E-03	6.53E-03	6.86E-03
Estimate effect size	**300**	8.85E-03	9.00E-03	9.33E-03	9.57E-03	1.77E-03	1.55E-03	1.09E-03	8.23E-04	3.55E-03	3.74E-03	3.02E-03	2.56E-03
	500	7.28E-03	8.16E-03	7.60E-03	8.13E-03	1.07E-03	1.10E-03	8.38E-04	4.46E-04	2.70E-03	2.73E-03	2.46E-03	2.21E-03
	1000	6.54E-03	6.92E-03	7.14E-03	7.25E-03	6.68E-04	6.17E-04	5.18E-04	3.30E-04	2.14E-03	2.11E-03	1.97E-03	1.81E-03
	5000	5.88E-03	6.10E-03	6.53E-03	6.86E-03	3.35E-04	2.99E-04	2.70E-04	2.25E-04	1.62E-03	1.67E-03	1.65E-03	1.64E-03

(b)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	-1.8E-02	-1.6E-02	-1.2E-02	-1.1E-02	-3.7E-02	-3.9E-02	-3.8E-02	-3.1E-02	-7.8E-02	-8.1E-02	-8.4E-02	-8.5E-02
	500	-1.6E-02	-1.5E-02	-1.5E-02	-1.0E-02	-3.6E-02	-3.7E-02	-3.7E-02	-3.4E-02	-7.5E-02	-8.1E-02	-7.9E-02	-8.2E-02
	1000	-1.5E-02	-1.2E-02	-1.4E-02	-1.1E-02	-3.6E-02	-3.5E-02	-3.4E-02	-3.5E-02	-7.6E-02	-7.8E-02	-8.1E-02	-8.2E-02
	5000	-1.4E-02	-1.4E-02	-1.2E-02	-1.1E-02	-3.6E-02	-3.5E-02	-3.5E-02	-3.5E-02	-7.6E-02	-7.7E-02	-8.0E-02	-8.2E-02
Estimate effect size	**300**	-1.9E-02	-1.8E-02	-1.5E-02	-1.7E-02	-4.2E-02	-4.5E-02	-4.1E-02	-3.9E-02	-8.4E-02	-8.8E-02	-9.2E-02	-9.2E-02
	500	-1.7E-02	-1.8E-02	-1.6E-02	-1.4E-02	-4.0E-02	-4.3E-02	-4.2E-02	-4.1E-02	-7.9E-02	-8.4E-02	-8.9E-02	-9.1E-02
	1000	-1.5E-02	-1.6E-02	-1.6E-02	-1.4E-02	-4.0E-02	-4.0E-02	-3.9E-02	-3.9E-02	-8.0E-02	-8.3E-02	-8.6E-02	-8.8E-02
	5000	-1.5E-02	-1.5E-02	-1.5E-02	-1.4E-02	-3.9E-02	-4.0E-02	-4.0E-02	-4.0E-02	-8.1E-02	-8.4E-02	-8.7E-02	-9.0E-02

Supplementary Table S2.4 (a-d): MSE of estimated specificity when using different number of variants. (a) 51 variants, (b) 25 variants, (c) 1 variant with the largest effect size (d)1 variant with moderate effect size

(a)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	9.42E-04	7.05E-04	4.22E-04	4.56E-05	1.30E-03	9.62E-04	6.84E-04	9.24E-05	2.12E-03	1.83E-03	1.00E-03	2.51E-04
	500	5.61E-04	3.77E-04	2.45E-04	1.86E-05	7.45E-04	5.85E-04	3.78E-04	5.08E-05	1.17E-03	1.03E-03	6.65E-04	1.21E-04
	1000	2.47E-04	1.95E-04	1.14E-04	8.52E-06	3.62E-04	2.87E-04	1.86E-04	2.36E-05	5.89E-04	4.90E-04	3.40E-04	5.95E-05
	5000	5.09E-05	3.71E-05	2.46E-05	1.50E-06	7.49E-05	5.61E-05	3.78E-05	3.92E-06	1.25E-04	9.11E-05	7.21E-05	1.24E-05
Estimate effect size	**300**	8.74E-04	6.58E-04	4.18E-04	6.17E-05	1.35E-03	8.91E-04	6.76E-04	1.49E-04	1.93E-03	1.65E-03	9.87E-04	4.15E-04
	500	5.15E-04	3.88E-04	2.58E-04	3.24E-05	7.61E-04	5.67E-04	3.63E-04	9.59E-05	1.11E-03	1.06E-03	6.94E-04	2.50E-04
	1000	2.62E-04	2.00E-04	1.33E-04	1.74E-05	3.81E-04	3.06E-04	2.01E-04	5.10E-05	6.19E-04	5.57E-04	4.22E-04	1.55E-04
	5000	6.38E-05	4.57E-05	3.16E-05	4.66E-06	9.95E-05	8.24E-05	6.06E-05	1.68E-05	2.05E-04	1.70E-04	1.43E-04	6.07E-05

(b)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	1.27E-03	9.99E-04	6.30E-04	1.08E-04	2.06E-03	1.49E-03	1.08E-03	2.85E-04	3.23E-03	2.89E-03	1.85E-03	8.25E-04
	500	7.81E-04	5.85E-04	3.89E-04	5.29E-05	1.15E-03	8.61E-04	5.87E-04	1.67E-04	2.00E-03	1.66E-03	1.21E-03	5.10E-04
	1000	3.65E-04	2.77E-04	1.82E-04	2.36E-05	5.43E-04	4.75E-04	3.09E-04	7.94E-05	9.91E-04	9.23E-04	6.49E-04	3.22E-04
	5000	7.92E-05	6.12E-05	4.12E-05	4.94E-06	1.40E-04	1.17E-04	8.32E-05	2.38E-05	3.35E-04	2.66E-04	2.47E-04	1.20E-04
Estimate effect size	**300**	1.22E-03	9.34E-04	6.46E-04	1.17E-04	1.96E-03	1.52E-03	1.06E-03	3.28E-04	2.99E-03	2.87E-03	1.85E-03	1.29E-03
	500	7.19E-04	5.91E-04	3.75E-04	7.10E-05	1.13E-03	9.29E-04	5.92E-04	2.31E-04	2.00E-03	1.87E-03	1.38E-03	7.27E-04
	1000	3.61E-04	3.14E-04	1.91E-04	3.90E-05	5.76E-04	5.13E-04	3.77E-04	1.40E-04	1.26E-03	1.15E-03	9.94E-04	5.35E-04
	5000	9.60E-05	7.37E-05	5.46E-05	1.19E-05	2.16E-04	2.00E-04	1.51E-04	5.95E-05	6.51E-04	6.08E-04	5.13E-04	3.07E-04

(c)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	1.43E-03	1.21E-03	7.08E-04	1.23E-04	2.38E-03	1.89E-03	1.26E-03	3.81E-04	4.49E-03	3.40E-03	2.28E-03	1.28E-03
	500	8.44E-04	6.91E-04	4.44E-04	7.42E-05	1.29E-03	1.07E-03	7.72E-04	2.31E-04	2.55E-03	2.16E-03	1.64E-03	9.12E-04
	1000	3.90E-04	3.23E-04	2.21E-04	3.68E-05	7.11E-04	5.33E-04	3.99E-04	1.41E-04	1.48E-03	1.23E-03	9.66E-04	5.35E-04
	5000	8.94E-05	6.82E-05	4.57E-05	7.72E-06	1.82E-04	1.45E-04	1.20E-04	3.92E-05	5.20E-04	4.87E-04	3.84E-04	2.40E-04
Estimate effect size	**300**	1.37E-03	1.09E-03	6.99E-04	1.39E-04	2.17E-03	1.72E-03	1.20E-03	4.44E-04	3.90E-03	3.58E-03	2.37E-03	1.38E-03
	500	8.05E-04	6.48E-04	4.80E-04	1.00E-04	1.26E-03	1.05E-03	7.55E-04	2.40E-04	2.53E-03	2.29E-03	1.74E-03	9.97E-04
	1000	3.83E-04	3.35E-04	2.30E-04	3.82E-05	6.73E-04	5.82E-04	4.00E-04	1.62E-04	1.42E-03	1.42E-03	1.11E-03	6.49E-04
	5000	9.91E-05	7.59E-05	5.37E-05	1.12E-05	2.04E-04	1.78E-04	1.41E-04	5.93E-05	6.45E-04	5.99E-04	5.15E-04	3.44E-04

(d)

		Prevalence=4%				Prevalence=10%				Prevalence=15%			
		0.85	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**	**0.85**	**0.9**	**0.95**	**1**
True effect size	**300**	6.43E-03	3.66E-03	1.35E-03	1.63E-04	7.39E-03	4.28E-03	2.00E-03	5.92E-04	9.91E-03	6.57E-03	3.41E-03	1.83E-03
	500	5.62E-03	2.73E-03	1.13E-03	1.06E-04	5.96E-03	3.50E-03	1.45E-03	3.30E-04	7.87E-03	4.89E-03	2.61E-03	9.62E-04
	1000	3.22E-03	1.73E-03	7.33E-04	5.46E-05	3.53E-03	2.26E-03	1.00E-03	1.80E-04	5.26E-03	3.48E-03	1.67E-03	5.48E-04
	5000	4.70E-04	3.10E-04	1.80E-04	7.18E-06	6.20E-04	4.84E-04	3.31E-04	2.45E-05	1.08E-03	9.07E-04	5.98E-04	8.30E-05
Estimate effect size	**300**	6.98E-03	3.87E-03	1.41E-03	1.64E-04	8.71E-03	4.56E-03	2.11E-03	5.83E-04	1.06E-02	6.71E-03	3.43E-03	1.65E-03
	500	6.13E-03	2.93E-03	1.18E-03	1.07E-04	6.92E-03	3.74E-03	1.53E-03	3.15E-04	8.75E-03	5.43E-03	2.49E-03	9.19E-04
	1000	4.49E-03	2.20E-03	7.93E-04	5.31E-05	4.17E-03	2.78E-03	1.14E-03	1.49E-04	6.14E-03	3.78E-03	1.75E-03	4.91E-04
	5000	8.40E-04	4.28E-04	2.17E-04	8.53E-06	9.25E-04	6.33E-04	4.31E-04	2.21E-05	1.55E-03	1.26E-03	7.76E-04	5.87E-05

Supplementary Table S2.5 (a,b): (a) RMSE and (b) bias for estimation of specificity when the EHR study uses a different liability threshold compared to external GWAS studies.

(a)

sample size	1000				5000			
delta / alpha_0	**-0.05**	**0**	**0.05**	**0.1**	**-0.05**	**0**	**0.05**	**0.1**
0.85	0.016	0.020	0.027	0.037	0.008	0.012	0.021	0.032
0.9	0.015	0.018	0.025	0.033	0.007	0.011	0.020	0.032
0.95	0.011	0.016	0.024	0.033	0.005	0.010	0.019	0.030
1	0.003	0.009	0.017	0.028	0.001	0.006	0.014	0.026

*delta: $K_{EHR} - K_{EX}$

(b)

sample size	1000				5000			
delta / alpha_0	**-0.05**	**0**	**0.05**	**0.1**	**-0.05**	**0**	**0.05**	**0.1**
0.85	-0.001	-0.005	-0.014	-0.024	0.000	-0.008	-0.018	-0.029
0.9	0.000	-0.005	-0.013	-0.021	0.000	-0.008	-0.018	-0.030
0.95	0.001	-0.005	-0.015	-0.022	0.000	-0.007	-0.017	-0.029
1	-0.001	-0.005	-0.012	-0.021	-0.001	-0.004	-0.013	-0.025

*delta: $K_{EHR} - K_{EX}$

Supplementary Table S2.6: Empirical type I error for testing whether there is misspecification by incorporating effect sizes estimated based on 20,000 samples.

	alpha=0.05				alpha=0.01			
Size / prevalence	**300**	**500**	**1000**	**5000**	**300**	**500**	**1000**	**5000**
4%	0.05	0.06	0.04	0.06	0.01	0.01	0.01	0.02
10%	0.05	0.07	0.06	0.08	0.02	0.01	0.02	0.03
15%	0.06	0.07	0.09	0.11	0.02	0.02	0.01	0.04

Supplementary Table S2.7 (a,b): Power for testing whether there is misspecification under different settings. (a) alpha = 0.05, (b) alpha = 0.01.

(a) alpha = 0.05

		Prevalence=4%			Prevalence=10%			Prevalence=15%		
		0.85	**0.9**	**0.95**	**0.85**	**0.9**	**0.95**	**0.85**	**0.9**	**0.95**
True effect size	**300**	1.00	1.00	0.86	1.00	0.96	0.72	0.96	0.84	**0.53**
	500	1.00	1.00	0.97	1.00	1.00	0.90	1.00	0.97	0.69
	1000	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.91
	5000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Estimate effect size	**300**	1.00	0.99	0.92	1.00	0.97	0.83	0.97	0.90	0.64
	500	1.00	1.00	0.99	1.00	1.00	0.94	1.00	0.98	0.81
	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
	5000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

(b) alpha = 0.01

		Prevalence=4%			Prevalence=10%			Prevalence=15%		
		0.85	**0.9**	**0.95**	**0.85**	**0.9**	**0.95**	**0.85**	**0.9**	**0.95**
True effect size	**300**	0.99	0.97	0.72	0.98	0.89	**0.49**	0.88	0.68	**0.29**
	500	1.00	1.00	0.92	1.00	0.98	0.74	0.98	0.90	**0.49**
	1000	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.99	0.79
	5000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Estimate effect size	**300**	1.00	0.98	0.80	0.99	0.93	0.64	0.91	0.77	0.39
	500	1.00	1.00	0.96	1.00	0.99	0.85	0.99	0.95	0.61
	1000	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.92
	5000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

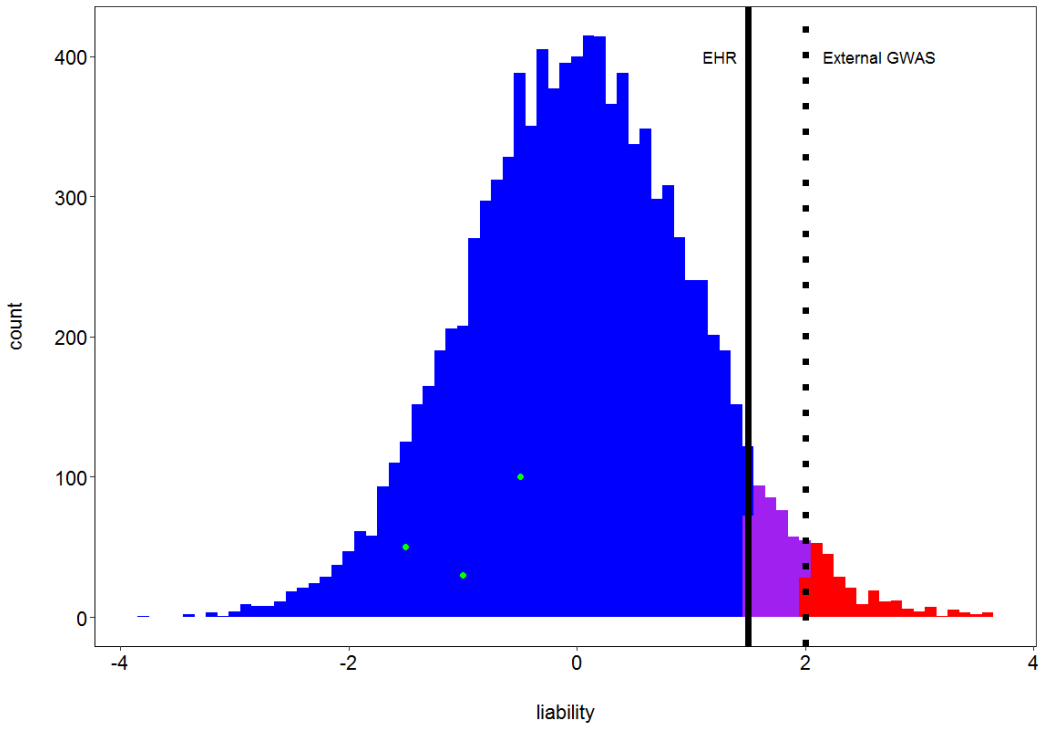
Supplementary Table S2.8: Estimated misclassification rate in observed cases and p-value when using different encounter cutoffs to define an AMD case. Holm-Bonferroni method is used to control the familywise error rates under 0.05. Numbers in bold represents the results that are not significant.

Cutoff	Cases	Controls	Estimated (1-α_0)	p-value	Holm-Bonferroni threshold*
≥1 encounter	144	16516	0.40%	1.94E-14	3.85E-03
≥2 encounters	119	16516	0.30%	2.10E-10	4.17E-03
≥3 encounters	86	16516	0.16%	7.85E-06	4.55E-03
≥4 encounters	74	16516	0.12%	2.35E-04	5.00E-03
≥5 encounters	65	16516	0.10%	7.37E-04	5.56E-03
≥6 encounters	55	16516	0.10%	1.44E-03	6.25E-03
≥7 encounters	49	16516	0.06%	**1.51E-02**	7.14E-03
≥8 encounters	45	16516	0.04%	**9.60E-02**	8.33E-03
≥9 encounters	44	16516	0.04%	**9.37E-02**	1.00E-02
≥10 encounters	42	16516	0.04%	**9.05E-02**	1.25E-02
≥11 encounters	38	16516	0.02%	**3.72E-01**	1.67E-02
≥12 encounters	37	16516	0.02%	**3.74E-01**	2.50E-02
≥13 encounters	33	16516	0.02%	**4.38E-01**	5.00E-02

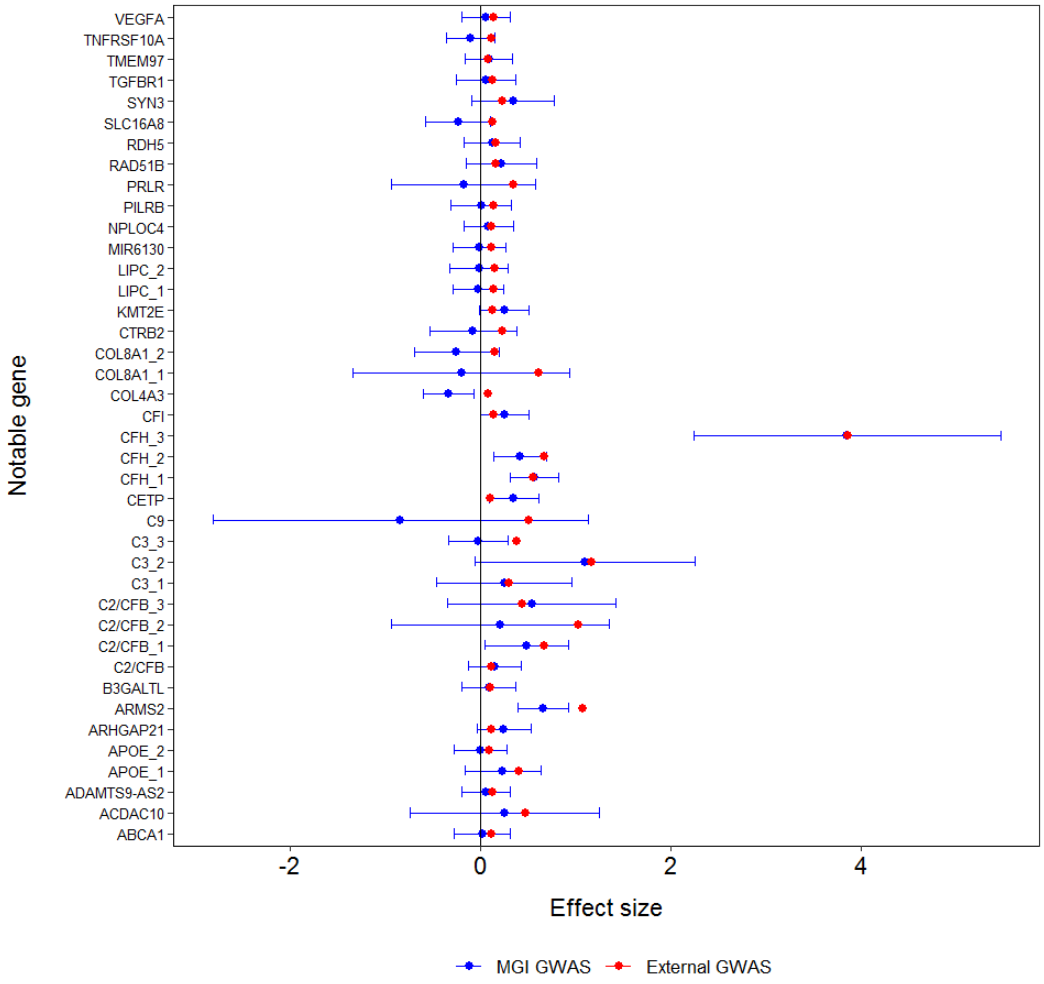
*Holm-Bonferroni threshold = 0.05/(13 – rank of the test + 1)

Supplementary Table S2.9: The estimation of MAFs in cases and effect sizes for AMD with samples defined by different encounter cutoffs. Here the top two common variants with largest effect sizes are investigated. Results are compared with those in the external GWAS to examine whether they are converge to the external gold-standard GWAS when phenotypes are refined by more stringent cutoff. The variant, rs3750846, is around gene *ARMS2*; and rs10922109 is around gene *CFH*.

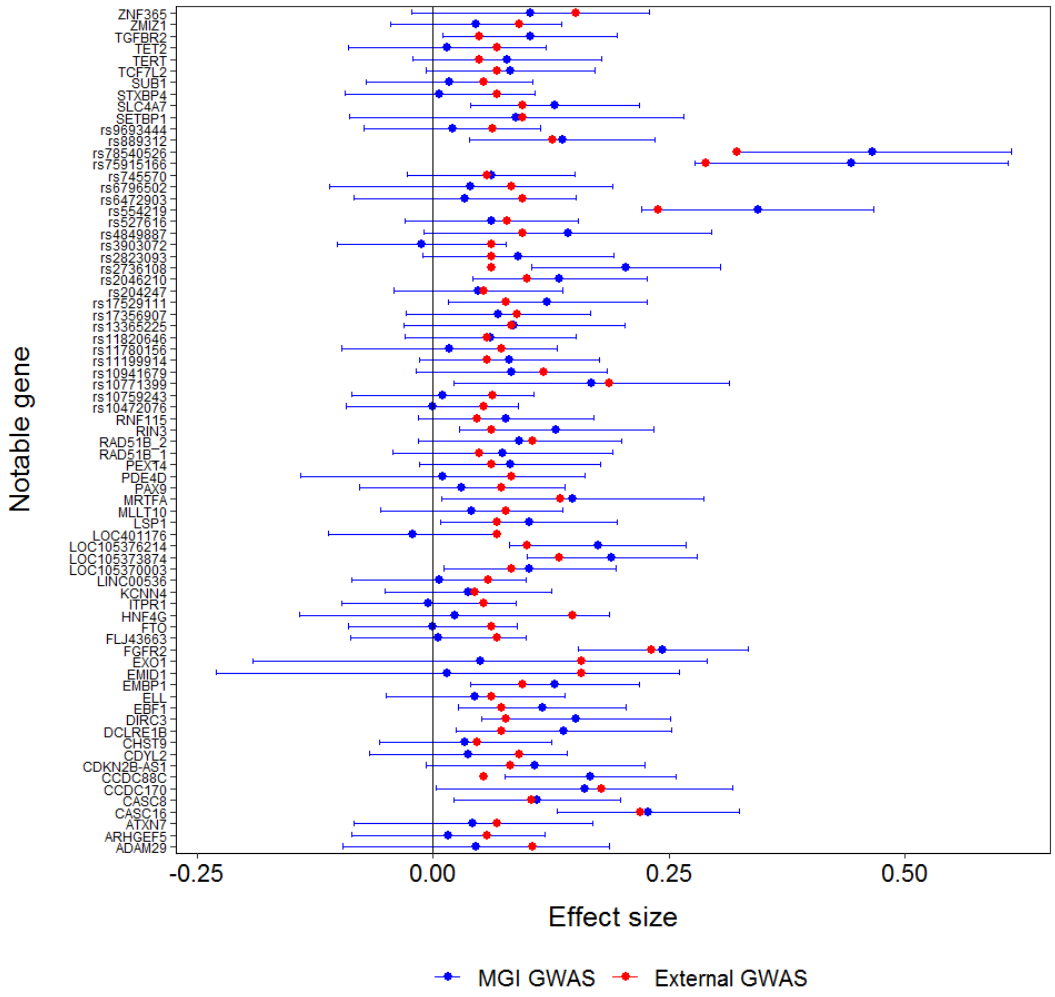
	MAF in cases		**Effect size**	
	rs3750846	**rs10922109**	**rs3750846**	**rs10922109**
≥1 encounter	0.347	0.323	0.634	0.382
≥2 encounters	0.353	0.315	0.660	0.417
≥3 encounters	0.384	0.302	0.793	0.477
≥4 encounters	0.412	0.291	0.912	0.534
≥5 encounters	0.415	0.277	0.925	0.601
≥6 encounters	0.400	0.273	0.861	0.622
≥7 encounters	0.418	0.255	0.938	0.713
≥8 encounters	0.444	0.233	1.045	0.831
≥9 encounters	0.455	0.239	1.086	0.802
≥10 encounters	0.464	0.250	1.126	0.740
≥11 encounters	0.487	0.237	1.217	0.812
≥12 encounters	0.500	0.243	1.270	0.777
≥13 encounters	0.515	0.242	1.330	0.782
External GWAS	**0.436**	**0.223**	**1.075**	**0.673**



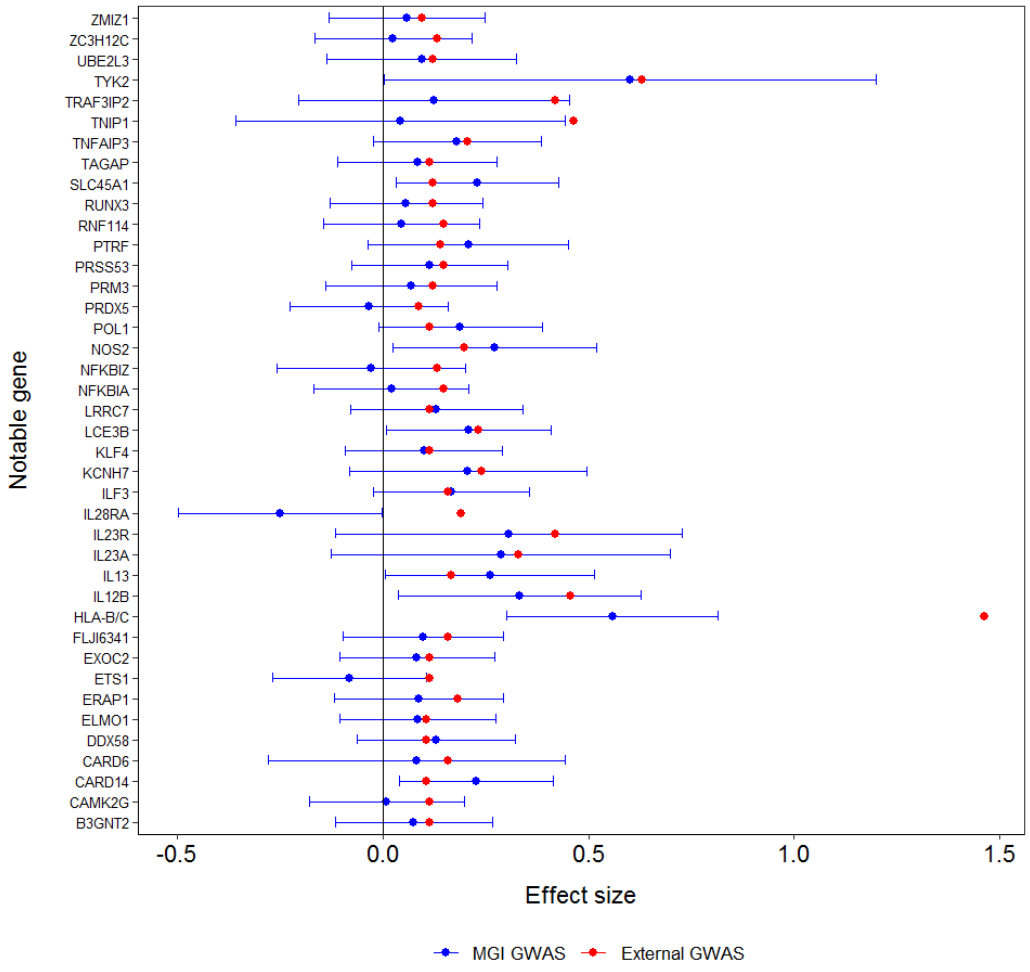
Supplementary Figure S2.1: Illustration of the influence of different case/control dichotomization thresholds on case/control distributions in external GWAS and EHR-based GWAS. Here EHR has less stringent liability threshold to dichotomize cases than external GWAS. The blue area represents samples that are classified as controls in both GWAS. The red area represents samples that are classified as cases in both GWAS. The green dots represent the samples that are truly misclassified. The purple area represents samples that are classified as cases in EHR and controls in external GWAS due to the difference in the liability thresholds.



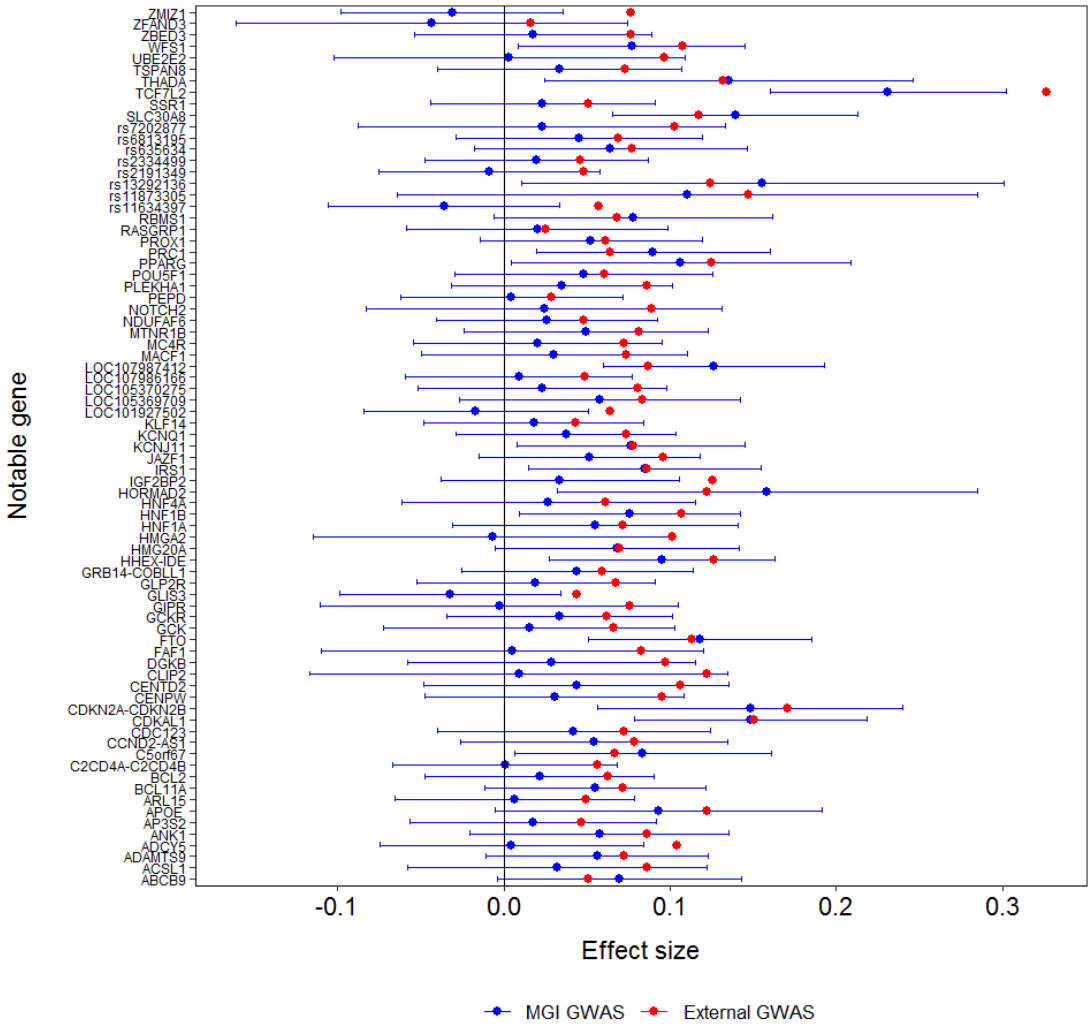
Supplementary Figure S2.2(a): Estimated effect sizes in external case-control study and EHR GWAS for age-related macular degeneration. Forty variants listed here are reported to be significantly associated with AMD by the International AMD Genomics Consortium and have association results in MGI EHR (Fritsche et al., 2016). On the y axis, if multiple variants are around the same notable gene, gene_* are used to distinguish them.



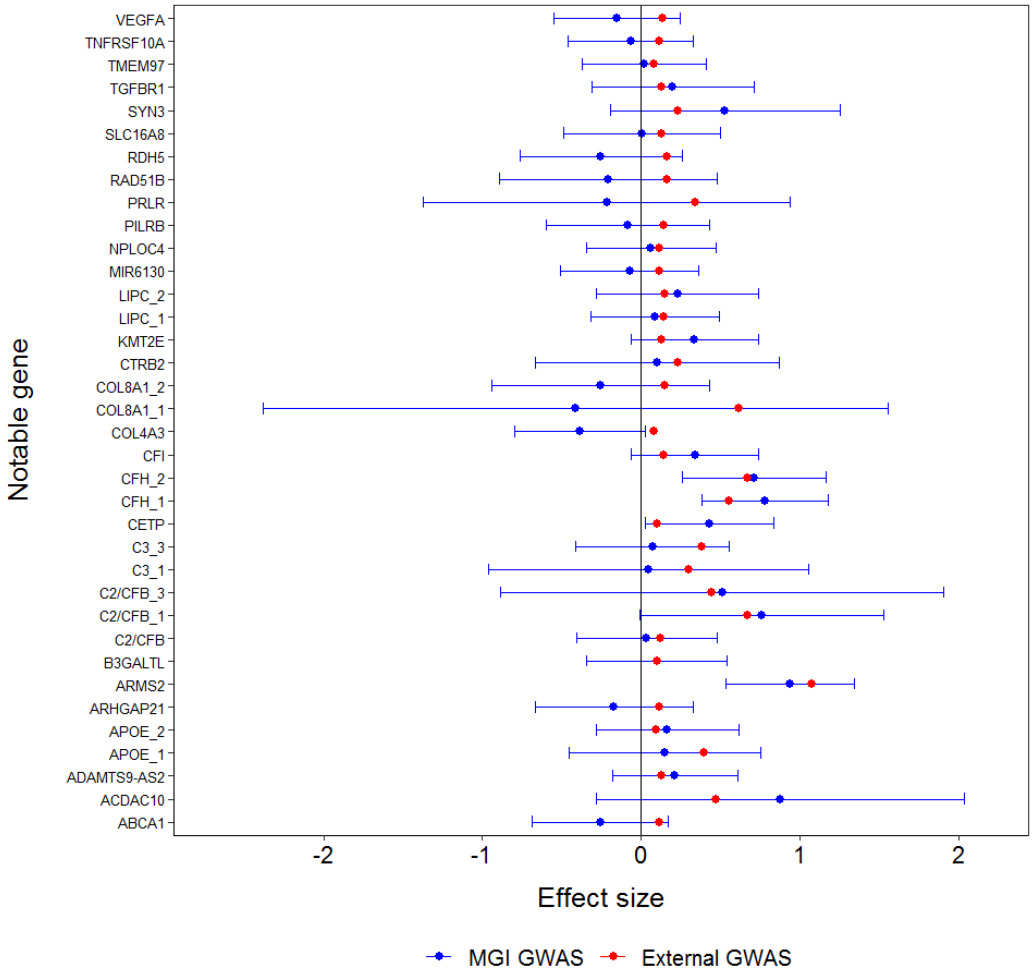
Supplementary Figure S2.2(b): Estimated effect sizes in external case-control study and EHR GWAS for breast cancer. Seventy three variants listed here are reported to be significantly associated with breast cancer in external studies (Michailidou et al., 2017; Michailidou et al., 2015; Michailidou et al., 2013; Fletcher et al., 2011; Turnbull et al., 2010) and have association results in MGI EHR. The external GWAS results are the average of effect sizes among those five studies. On the y axis, if multiple variants are around the same notable gene, gene_* are used to distinguish them. If no gene is around the variant, its rs ID is shown instead.



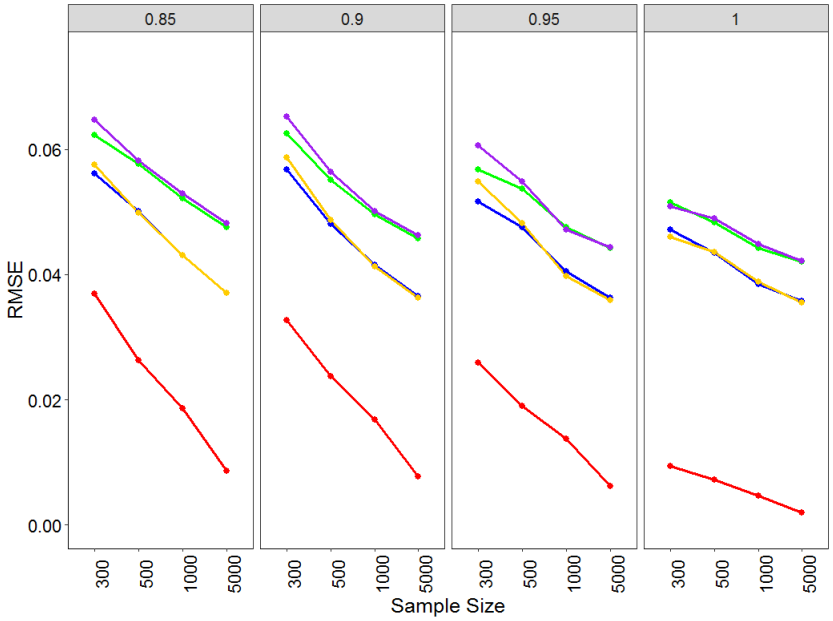
Supplementary Figure S2.2(c): Estimated effect sizes in external case-control study and EHR GWAS for psoriasis. Forty one variants listed here are reported to be significantly associated with psoriasis in external studies (Tsoi et al., 2015; Yin et al., 2015; Tsoi et al., 2012; Strange et al., 2010; Stuart et al., 2010) and have association results in MGI EHR. The external GWAS results are the average of effect sizes among those five studies.



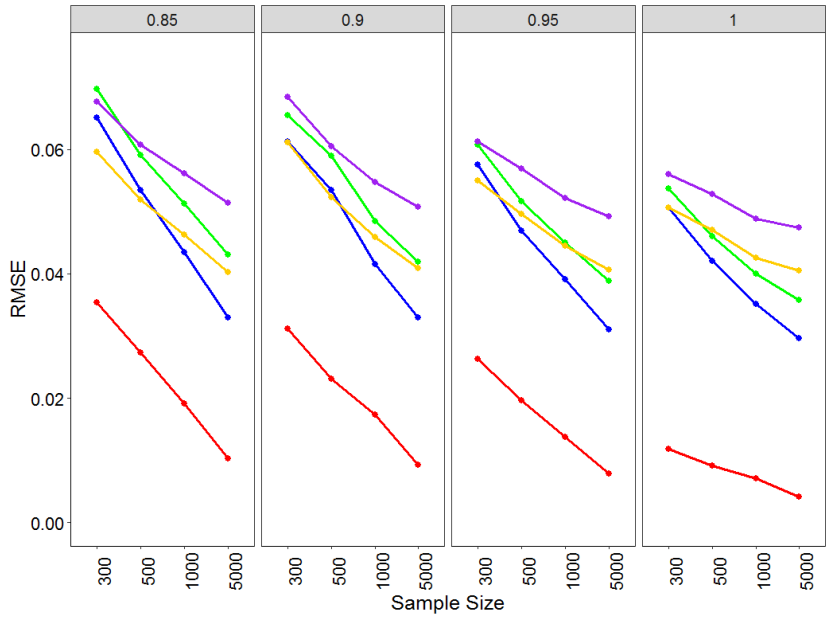
Supplementary Figure S2.2(d): Estimated effect sizes in external case-control study and EHR GWAS for type II diabetes. Seventy seven variants listed here are reported to be significantly associated with T2D in external studies (Scott et al., 2017; Gaulton et al., 2015; Consortium, Diabetes SAT2D, et al., 2014; Morris et al., 2012; Saxena et al., 2012; Voight et al., 2010; Zeggini et al., 2008) and have association results in MGI EHR. The external GWAS results are the average of effect sizes among those seven studies. On the y axis, if no gene is around the variant, its rs ID is shown instead.



Supplementary Figure S2.3: Estimated effect sizes in external case-control study and EHR GWAS using samples having 7 or more encounters for age-related macular degeneration. Thirty six variants listed here are reported to be significantly associated with AMD by the International AMD Genomics Consortium and have converged association results in MGI EHR (Fritsche et al., 2016). On the y axis, if multiple variants are around the same notable gene, gene_* are used to distinguish them.



(a)



(b)

Supplementary Figure S2.4(a,b): RMSE for estimation of specificity based on different methods; (a) is using true effect sizes and (b) is using estimated effect sizes. Results shown here are for settings with disease prevalence 10%. The panel represents true specificity α_0. The red line represents results using our method. The blue line and yellow line represent method proposed by Tsoi et al. that examines mean differences with exact risk AF and approximate risk AF respectively. The green line and purple line represent their method that examines median differences with exact risk AF and approximate risk AF respectively.

CHAPTER III

Likelihood-based Protocol for Inferring Genetic Relatives Securely between Studies

3.1 Introduction

The development of next generation sequencing technologies has benefitted many areas of genetic research. Based on whole genome or exome sequencing data, reference datasets of human genetic variation have been generated (Karczewski et al., 2019; Taliun et al., 2019). They are essential resources for functional interpretation of putative disease-causing variants by, for example, helping separate genomic positions and regions that are mutation intolerant from others where variation is more common. Electronic health records (EHR) has also been linked to sequencing data, for example, the exome sequencing project of UK Biobank samples funded by Regeneron Pharmaceuticals (UK Biobank, 2018). These sequencing data from large cohorts can be used to establish the reference dataset. As more and more reference resources like dbSNP and ClinVar, or custom browsers like gnomAD and BRAVO, became publicly available, aggregated information from multiple resources will help researchers make inferences more efficiently and comprehensively (Sherry et al., 2001; Landrum et al., 2017; Karczewski et al., 2019; Taliun et al., 2019). The critical step of the combination of different reference datasets is to infer the overlapping samples, as ignoring overlapping samples when combining information will bias the

summary statistics like allele frequency (AF), especially for rare variants. In turn, ignoring overlap will be deleterious for functional interpretation of these variants.

Various methods have been developed to infer genetic relatives based on individual-level genetic data (Lynch, 1988; Queller and Goodnight, 1989; Boehnke and Cox, 1997; Broman and Weber, 1998; Lynch and Ritland, 1999; Epstein et al. 2000; Wang, 2002; Milligan, 2003; Manichaikul, 2010; Thomas, 2010). However, methods are lacking for a more challenging problem, inferring genetic relatives between different studies. An inherent issue that arises in identifying overlapping or closely related samples between studies is privacy. While the summary statistics are often shared between studies, it is common that studies are prohibited from sharing individual-level data most often because of the informed consent used in the studies. Moreover, as more and more genetic data are gathered and stored in large databases, they may become a resource for people to find their genetic relatives. Privacy-preserving protocols of inferring relatives can also be applied when people who are interested in finding their relatives have concerns about releasing genetic data to organizations they may not necessarily trust.

Several methods initially aiming for secure DNA string searching or edit distance calculation have been proposed based on different cryptographic techniques. An extension usage of these two-party secure protocols is to identify similar samples, i.e. infer duplicates, between two studies based on the edit distance. The protocols include private set intersection (De Cristofaro and Tsudik, 2012), oblivious transfer- based hamming distance system (Bringer et al., 2013), as well as privacy-preserving approximating edit distance (Wang et al., 2015). These methods can infer duplicates by calculating the similarity or distance of encrypted genome sequences. However, all of these methods are not constructed based on the biological mechanism of inheritance and can only infer duplicates. Another type of method deals with this problem based

on the ‘fuzzy’ encryption technique. He et al. (2014) showed that with this ‘fuzzy’ encryption method, one individual should be able to decrypt the encrypted genome of another individual using his own genome only when they are related. The extension of the method proposed by Hormozdiari et al (2014) takes advantage of genetic reference panels and can detect more distant relationships using rare variants. However, they also have some limitations. One limitation is that they use the haplotype information, which requires phasing of the genotype. Second, these methods require sharing information of the whole genome between studies. As a consequence, they are not only computationally infeasible for large-scale data, but also have a risk of severe information leakage when the encryption is attacked. More importantly, these methods, again, are based on comparing the similarity of genome sequences directly, but ignoring the mechanism of inheritance. In this chapter, we propose a new protocol that allows efficient detection of genetic relatives without compromising privacy while only requiring sharing encrypted information of a limited number of variants.

The secure protocol infers the genetic relationship using a likelihood-based model. This method was first proposed by Boehnke and Cox in 1997 and improved by Broman and Weber (1998) and Epstein et al. (2000) to incorporate genotyping error. The likelihood-based method has excellent power of identifying genetic relatives using true individual-level genotype data. Here, we modify this method and enable it to identify relatives using encrypted genotype data. The general framework of this protocol between two studies is that study A first releases the encrypted code of their genotype segments to study B. Then study B, which has access to its own genotype data and the encrypted data from study A, can identify relatives by incorporating the likelihood-based method. Under this protocol, kinship coefficients can be obtained without disclosing genetic information between studies. We demonstrate the utility of our method by applying our

technique to infer genetic relatives among samples from the Trans-Omics for Precision Medicine (TOPMed) study. We show that our protocol can infer relationship up to second degree in a homogenous population with only a limited number of variants. Furthermore, with selected variants, our protocol can identify close relatives in a heterogeneous population. The computation time of our protocol scales well in practice. Our method is not limited to EHR studies, but can be broadly applied to any genetic studies without additional assumptions.

3.2 Method

3.2.1 Likelihood-based method of inferring genetic relatives

To infer relationship based on likelihoods, we have to construct the probability of observing the genotype pairs of two samples given a certain relationship (Boehnke and Cox, 1997; Epstein et al., 2000). Then the relationship which maximizes the likelihood is inferred to be the relationship between these two samples.

Let $G_m = (G_{im}, G_{jm})$ denote the genotype of sample pair (i, j) at variant m. If M independent variants are selected to infer relationship, $G = (G_1, G_2, \dots, G_M)$ should be the genotype pairs of all the M variants for a pair of samples i and j. Then we can infer the relationship based on the probability of observing the genotype pairs between the samples given a certain relationship, $Pr(G|R)$. The relationship we consider here includes MZ twins/duplicates, parent-offspring, full siblings, second degree relatives and unrelated pairs. For example, if $Pr(G|full\ sib)$ is the largest likelihood among all relationships, it suggests that this pair of samples may be full siblings.

To calculate $Pr(G|R)$, we use the identity-by-descent (IBD) status to link the observed genotype pairs G and relationship R. Let $IBD_m \in \{0,1,2\}$ denote the number of alleles shared by IBD at variant m. Then for a pair of variants G_m of sample i and j, the probability can be calculated as

$$Pr(G_m|R) = \sum_{d=\{0,1,2\}} Pr(G_m|IBD_m = d) \cdot Pr(IBD_m = d|\, R). \quad (1)$$

The first probability in the equation, $\Pr(\mathrm{G_m}|\mathrm{IBD} = \mathrm{i})$, is the conditional probability of genotype pairs at variant m given that they share 0,1 or 2 alleles IBD. These probabilities are provided in Table 3.1 for autosomal variants (Thompson, 1975; Risch, 1990). Here we only consider bi-allelic variants.

Table 3.1: Probability of ordered autosomal genotype pairs given IBD status $\Pr(\boldsymbol{G_m}|\boldsymbol{IBD_m} = \mathbf{0,1,2})$.

	Pr(Genotype pairs|IBD)		
Genotype pairs	**IBD=0**	**IBD=1**	**IBD=2**
(aa,aa)	p_a^4	p_a^3	p_a^2
(aa,ab)	$2p_a^3p_b$	$p_a^2p_b$	0
(aa,bb)	$p_a^2p_b^2$	0	0
(ab,ab)	$4p_a^2p_b^2$	$p_a^2p_b + p_ap_b^2$	$2p_ap_b$

Given a relationship, we can construct the probability of having IBD status to be 0, 1 or 2 at that variant for autosomal variants. For instance, $Pr(IBD_m = 0,1,2|Full\ sibs) = (0.25,0.5,0.25)$. Table 3.2 shows these probabilities for different relationships.

Table 3.2: Probability of IBD status given different relationship $\Pr(\boldsymbol{IBD_m} = \mathbf{0,1,2}|\boldsymbol{R})$.

	Pr(IBD|R)		
Relationship	**IBD=0**	**IBD=1**	**IBD=2**
Duplicates/MZ twins	0	0	1
Parent-offspring	0	1	0
Full-sib	0.25	0.5	0.25
Second-degree relatives	0.5	0.5	0
First-cousin	0.75	0.25	0
Unrelated	1	0	0

For our protocol, we only select independent variants to infer the relationship. Thus, the joint log-likelihood of the genotype pairs of all M variants, given the relationship, should be the log of the product of these independent variants,

$$l(G|R) = log\left(\prod_{m=1}^{M} Pr(G_m|R)\right)$$

$$= \sum_{m=1}^{M} log\left(\left(\sum_{d=\{0,1,2\}} Pr(G_m|IBD_m = d) \cdot Pr(IBD_m = d|R)\right)\right). \quad (2)$$

3.2.2 Genotyping error

Broman and Weber (1998) demonstrated that if we did not consider genotype errors in the model, there would be many probabilities of value 0 introduced in the model through $Pr(G_m|IBD_m)$, which reduced the accuracy and flexibility of the model. For example, the probability $Pr(G_m = (aa, bb)|IBD_m = 1)$ equals 0 when we ignore genotype errors. However, it is possible to get the genotype pair (aa,bb) under IBD = 1 when genotype error is considered. To make the model more realistic, they proposed a method considering genotype errors. Suppose each variant genotype is wrong with probability ϵ and each genotype is correctly determined with probability 1- ϵ. The probabilities $Pr(G_m|IBD_m)$ becomes the weighted sum of $Pr(G_m|IBD_m)$ and $Pr(G_m|IBD_m = 0)$. The weights are the probabilities that a pair of variants is correctly genotyped, $(1 - \varepsilon)^2$, and either variant is randomly generated from the population, $(1 - (1 - \varepsilon)^2)$. Hence, $Pr(G_m|IBD_m)$ becomes

$$Pr(G_m|IBD_m = d; \varepsilon) = (1 - \varepsilon)^2 Pr(G_m|IBD_m = d; \varepsilon = 0)$$

$$+(1 - (1 - \varepsilon)^2)Pr(G_m|IBD_m = 0; \varepsilon = 0). \quad (3)$$

This equation (3) is plugged into (2) to construct a likelihood which takes genotype error into consideration,

$$l(G|R,\varepsilon) = \sum_{m=1}^{M} log\left(\left(\sum_{d=\{0,1,2\}} Pr(G_m|IBD_m = d;\varepsilon)\cdot Pr(IBD_m = d|R)\right)\right).$$

3.2.3 General secure relationship inference framework

In the previous section, we demonstrated how to infer genetics relatives using individual-level genotypes. To infer relationships between studies when the individual-level genotype is not sharable, we need to calculate the likelihood defined above using encrypted genotype data. We propose a privacy-preserving framework that allows two studies to infer genetic relatives without exposing their individual-level genotype information. Two parties included in this protocol, study A and B, should follow these general steps (illustrated in Figure 3.1):

1) Study A generates encrypted genotype data and sends it to study B.
2) Study B calculates the likelihood using encrypted genotype from study A and its own genotype. Then study B infers the relationship based on the likelihoods.

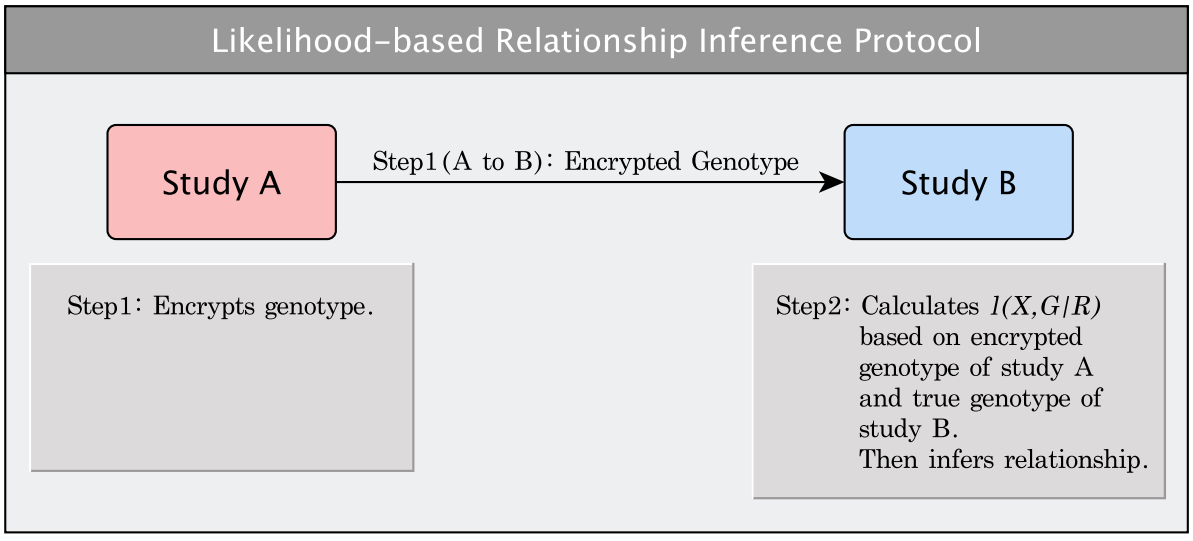


Figure 3.1: General framework for securely calculating kinship coefficient between studies.

3.2.4 Inferring relationships using encrypted genotype by likelihood-based method

As described in the previous section, to infer relationships securely, the encryption scheme should be developed and the likelihood $l(G|R,\varepsilon)$ depending on encrypted genotype should be constructed.

The key concept of encrypting genotypes is to represent true genotype data with summary statistics that do not reveal the genotype specifically for each variant, but contains enough information to infer relationships. To encrypt genotype data, we first partition an individual's genotype into segments. Each segment contains k number of variants. Then, we summarize the genotypes within each segment as the encrypted code and use them to infer relatives. Let X_{ip} denote the encrypted genotype code for individual i at segment p, $X_{ip} = \sum_{m\in\{segment\ p\}} G_{im}$. One example of the encryption scheme with segment size k=3 is shown in Figure 3.2 and Table 3.3. In this encryption scheme, since the mappings between the encrypted genotype and the true genotype segment are one-to-one for (X = 0, G = (000)) and (X = 6, G = (222)), we set their encrypted genotype code to be 1and 5 instead of 0 and 6. In turn, the mappings between encrypted genotype and true genotype segment are always one-to-N. The information of the genotype segment is compressed from 3^3 distinct values into 5 distinct values. In other words, we cannot infer the exact true genotype from the encrypted genotype code. Details of the encryption scheme of segment size k = 5 are shown in Supplementary Table S3.1. With segment size equaling 5, the genotype is compressed from 243 distinct values to 9 distinct values. In Section 3.3 below, we will compare the utility and the security of these two encryption schemes through simulations.

Based on the likelihood constructed in Section 3.2.1-2 as well as the mappings between encrypted genotypes and true genotype segments, the joint log-likelihood of the pair of encrypted

Table 3.3: Mapping between true genotype and encrypted genotype code (segment size = 3).

Genotype segment (G)	Encrypted genotype code (X)
000, 001, 010, 100	1
011, 101, 110, 002, 020, 200	2
012, 021, 102, 111, 120, 201, 210	3
022, 112, 121, 202, 211, 220	4
122, 212, 221, 222	5

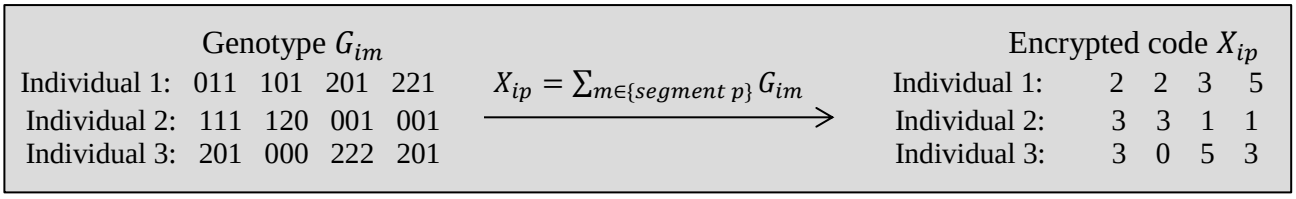


Figure 3.2: Demonstration of encrypting genotype with segment size = 3.

genotype from one study and the true genotype from the other study given relationship R is obtained by adding all possible conditions for a given encrypted genotype. Based on our protocol, for each comparison between sample i and j, genotypes of only one sample, sample i are encrypted. Suppose the sequence of genetic variants is divided into P segments. Each segment contains k variants. Let $\boldsymbol{G}_{ip}$ denote all the possible genotype segments that can be mapped to the encrypted genotype code X_{ip} in the pth segment, $X_{ip} = enc(\boldsymbol{G}_{ip}) = \sum_{m \in \{segment\ p\}} G_{im}$. Since each variant is independent, each segment should be independent also, and each variant within one segment should be independent as well, $Pr(\boldsymbol{G}_{ip}, \boldsymbol{G}_{jp} | R, \varepsilon) = \prod_{m \in \{segment\ p\}} Pr(G_m | R, \varepsilon)$. Hence, the likelihood becomes

$$l(\boldsymbol{X}_i, \boldsymbol{G}_j | R, \varepsilon) = \sum_p log(Pr(X_{ip}, \boldsymbol{G}_{jp} | R, \varepsilon))$$

$$= \sum_p log\left(\sum_{\boldsymbol{G_{ip}} \in \{enc(\boldsymbol{G_{ip}}) = X_{ip}\}} Pr(\boldsymbol{G}_{ip}, \boldsymbol{G}_{jp} | R, \varepsilon) \right)$$

$$= \sum_{p} log\left(\sum_{\boldsymbol{G_{ip}} \in \{enc(\boldsymbol{G_{ip}}) = X_{ip}\}} \prod_{m \in \{segment\ p\}} Pr(G_m | R, \varepsilon) \right).$$

The probability, $Pr(G_m|R, \varepsilon)$, can be calculated using equation (1) and (3). As in the method without encryption, the relationship is inferred by selecting the relationship which maximizes the likelihood.

To deal with missing values, if a genotype is missing for a sample in study A which encrypts the genotype, the whole segment corresponding to that variant is coded as missing and $Pr\left(X_{ip}, \boldsymbol{G}_{jp} \middle| R, \varepsilon\right)$ is set to be 1 for that segment. If the genotype is missing in study B which uses the true genotypes, we treat that variant as missing and $Pr(G_m|R, \varepsilon)$ is set to be 1 for that variant

3.3 Results

3.3.1 NHLBI TOPMed program

The performance of this secure relationship inference protocol is evaluated through application to infer relationships among the samples of NHLBI TOPMed program Freeze3 data.

TOPMed is a program that aims to get insight into the genetic basis of human diseases, including heart, lung, blood, and sleep disorders, through whole genome sequencing. The ethnic background of the participants in TOPMed program is diverse (Taliun et al., 2019). A reference dataset of human genetic variation, BRAVO, has been constructed based on the sequencing data providing resources for functional interpretation of the variants (Taliun et al., 2019). Inferring relationship securely among TOPMed samples and samples from other programs may allow us to aggregate information from multiple resources and develop a more comprehensive picture of the putative disease-causing variants.

3.3.2 Performance of the privacy preserving protocol in homogeneous populations

In the first simulation, we apply our method to infer relationships among homogeneous populations in TOPMed program. The ancestry of each individual is estimated using TRACE (Wang et al., 2015). Among 14,572 participants, we identify 3,357 Europeans, 3,437 Africans, 265 Asian and 54 Native Americans. Then relationships are inferred within each ethnic group. Gold-standard relationships among samples are inferred using the robust relationship inference method implemented in KING based on ~600,000 variants from the Human Genome Diversity Project (HGDP), which are considered to have high genotype quality (Cavalli-Sforza , 2005; Manichaikul et al. 2010). Then we conduct the secure inference of relationships using 500, 1000, 5000 and 10000 independent variants. These variants all have minor allele frequencies (MAF) around 0.5 (MAF: 0.4~0.5) in the joint population, which is representative of the most informative variants. We set the segment size to 3. Since the variants numbers are not a multiple of 3, the inference is actually conducted upon 498, 999, 4998, and 9999 variants.

To mimic the secure protocol, for each pair of samples, we assume the sample with smaller ID is from study A; it follows the protocol by encrypting its genotype of the selected variants. The sample with larger ID is from study B; it calculates the likelihood and infers the relationship based on encrypted data from study A and its own genotype. The allele frequencies we use for calculating the likelihood are based on all samples within each ethnic group.

We infer the relationships using our secure protocol with encrypted genotypes as well as using KING with the same set of unencrypted genotypes. The performance is evaluated by comparing the results of both methods to the gold standard. Here, considering that our final purpose of implementing the method is to identify overlaps between the 2 studies, our primary concern is identifying duplicates and 1st degree relatives.

Table 3.4 shows the number of relative pairs inferred totally and inferred correctly with different numbers of variants. With both methods, all of the duplicate pairs are correctly inferred using 500 or more variants.

For 1st degree relatives, the information we lose due to the encryption does not have a large impact on the relationship inference. Our method provides comparable results as KING without encryption. For instance, with both methods, we do not get false positives in the Asian and Native American populations. For the European and African populations, the number of correctly identified relative pairs increases and the false discovery rate (FDR) decreases as we use more variants. Our secure protocol recovers 100% of the 1st degree relatives and gets 0.15% false-positive pairs using 5000 variants for Europeans while KING without encryption identifies 100% of the 1st degree relatives with no false-positive pairs. For Africans, our method identifies 99.93% of the 1st degree relatives and gets 0.14% false-positive pairs using 5000 variants while KING without encryption identifies all the true pairs with 0.07% false-positive pairs.

In addition to 1st degree relatives, when using more than 5000 variants, 2nd degree relatives can also be identified with a high true positive rate and a low false discovery rate. The compression of information in the encrypted data has a large impact on the 2nd degree relative inference, which is mainly reflected in the false discovery rate. Using 10000 variants, our secure method detects 98.41% of the 2nd degree relatives in Europeans with 28.76% FDR while KING without encryption detects 95.66% of the 2nd degree relatives with 5.07% FDR. Our method detected 98.47% of the 2nd degree relatives with 3.28% FDR in Africans while KING without encryption detected 98.19% of the 2nd degree relatives with 1.72% FDR. Our method has identical performance as KING without encryption in Asian and Americans. Overall, in homogeneous

populations, our method does not suffer from the information loss due to the encryption and can infer relatives up to 2nd degree with high accuracy.

Table 3.4: Number of relative pairs inferred with 500, 1000, 5000 and 10000 variants within each ethnic group using our method (segment size =3) vs. KING without encryption. Data in the bracket represents number of correctly inferred pairs compared with gold-standard.

Method	Number of variants	European	African	Asian	American
Duplicates/MZ twins					
Gold standard		30	30	0	0
Our method	500	30(30)	30(30)	0	0
	1000	30(30)	30(30)	0	0
	5000	30(30)	30(30)	0	0
	10000	30(30)	30(30)	0	0
KING w/o encryption	500	30(30)	30(30)	0	0
	1000	30(30)	30(30)	0	0
	5000	30(30)	30(30)	0	0
	10000	30(30)	30(30)	0	0
1st degree relatives					
Gold standard		1959	1461	0	0
Our method	500	2184(1941)	1522(1448)	0	0
	1000	2070(1956)	1468(1453)	0	0
	5000	1962(1959)	1462(1460)	0	0
	10000	1960(1959)	1461(1461)	0	0
KING w/o encryption	500	1992(1941)	1475(1451)	0	0
	1000	1972(1951)	1461(1456)	0	0
	5000	1959(1959)	1462(1461)	0	0
	10000	1959(1959)	1461(1461)	0	0
2nd degree relatives					
Gold standard		2074	1049	1	0
Our method	**500**	68406(1560)	57825(779)	326(1)	5(0)
	1000	9629(1801)	3532(904)	16(1)	0
	5000	3045(2024)	1075(1014)	1(1)	0
	10000	2865(2041)	1068(1033)	1(1)	0
KING w/o encryption	**500**	17904(1549)	20306(889)	195(0)	2(0)
	1000	2714(1723)	1638(807)	6(1)	0
	5000	2088(1954)	1045(1020)	1(1)	0
	10000	2090(1984)	1048(1030)	1(1)	0

3.3.3 Performance of the privacy preserving protocol in a heterogeneous population

The calculation of the likelihood, $l(\boldsymbol{X}_i, \boldsymbol{G}_j | R, \varepsilon)$ depends on the allele frequencies of the variants, which in turn depend on the population background of the samples. In this section, we demonstrate the performance of the method in a heterogeneous population and provide a strategy about how to improve the performance of our method in this population.

In this simulation, relationships are identified among all the 7,113 participants in Section 3.3.2. The samples having diverse ethnic backgrounds like European, African, Asian and Native American are inferred together, assuming the ancestry information is not known. First we conduct the inference of relationships on the same set of independent 500, 1000, 5000 and 10000 variants that we used for the homogenous population application. The AF of some of these variants have significant discrepancies across different ancestries. The difference in ethnic-specific AF can be as large as 0.86 (0.11 for African vs. 0.98 for Asian). However, since we assume the participants are inferred without the ancestry information, the differences in AF are ignored. The allele frequencies we incorporate in the likelihood model are calculated based on the joint population.

Table 3.5 and Supplementary Table S3.2 show the number of relative pairs inferred totally and inferred correctly with different numbers of variants. All duplicate pairs are correctly inferred using 500 or more variants.

For 1st degree relatives, we are able to identify almost 100% of the related pairs with 1,000 or more variants. However, in terms of the false-positive pairs, we do suffer from incorporating the biased allele frequencies into the model; in other words, the false discovery rate is inflated. With 500 variants, about 70% of the inferred 1st degree relatives are not true 1st degree relatives. Even with 10000 variants, we still get 11.54% false-positive pairs using our method while the FDR is

0% using KING. This result implies that the inflated false positive results using our method are substantial when dealing with a heterogeneous population.

Table 3.5: Number of relative pairs inferred in a heterogeneous population using our method (segment size =3) with randomly selected variants vs. using KING without encryption. Data in the bracket represents number of correctly inferred pairs compared with gold-standard.

Method	Number of variants	Number of pairs inferred (correctly inferred)
Duplicates/MZ twins (Gold standard = 60)		
Our method w/ randomly selected variants	500	60(60)
	1000	60(60)
	5000	60(60)
	10000	60(60)
KING w/o encryption	500	60(60)
	1000	60(60)
	5000	60(60)
	10000	60(60)
1st degree relatives (Gold standard = 3425)		
Our method w/ randomly selected variants	500	11410(3414)
	1000	8532(3425)
	5000	4084(3425)
	10000	3872(3425)
KING w/o encryption	500	3472(3397)
	1000	3443(3412)
	5000	3426(3425)
	10000	3425(3425)

In order to tackle the inflated false-positive problem, we propose a variant selection strategy. Instead of selecting variants randomly, we select variants that have relatively constant allele frequencies across different ancestry groups. Variant selection is conducted based on two criteria: 1) differences of allele frequencies among 4 ancestries are less than 0.1; 2) differences of allele frequencies are less than 0.2. In our data, for all independent variants with allele frequency ranging from 0.4~0.5, only 789 variants are kept after being filtered by criterion-1. Thus here we do not infer the relationship using 1000 or 5000 variants under criterion-1.

The false discovery rate in detecting 1st degree relatives decreases significantly when using the selected variants compared with using randomly selected variants (Table 3.6, Figure 3.3). For instance, with 500 variants, the FDR decreases from 70.08% for randomly selected variants to 7.21% for criterion-1 based variants and 9.51% for criterion-2 based variants. When the false-positive result is reduced, the power to infer true related pairs remains the same (99.68% for random variants vs. 98.77% for criterion-1 vs. 98.92% for criterion-2). Overall, using more variants provides us with more accurate inference. With 5000 variants selected based on criterion-2, we are able to detect 99.94% of the 1st degree relatives with a 0.12% FDR. The results are almost identical to using KING with the unencrypted genotypes. However, while our protocol provides inference of duplicates and 1st degree relatives with high accuracy, the false discovery rate of detecting 2nd degree relatives is still not well controlled. The results of detecting 2nd degree relatives are shown in Supplementary Table S3.3. Unlike for homogenous populations, our protocol has a limited utility of inferring 2nd order relatives for heterogeneous populations.

Table 3.6: Number of relative pairs inferred in a heterogeneous population using our method (segment size =3) with variants selected based on different criteria. Data in the bracket represents number of correctly inferred pairs compared with gold-standard.

Method	Number of variants	Number of pairs inferred (correctly inferred)
Duplicates/MZ twins (Gold standard =60)		
Our method w/ criterion-1	500	60(60)
	800	60(60)
Our method w/ criterion-2	500	60(60)
	1000	60(60)
	5000	60(60)
1st degree relatives (Gold standard = 3425)		
Our method w/ criterion-1	500	3646(3383)
	800	3509(3401)
Our method w/ criterion-2	500	3744(3388)
	1000	3500(3412)
	5000	3427(3423)

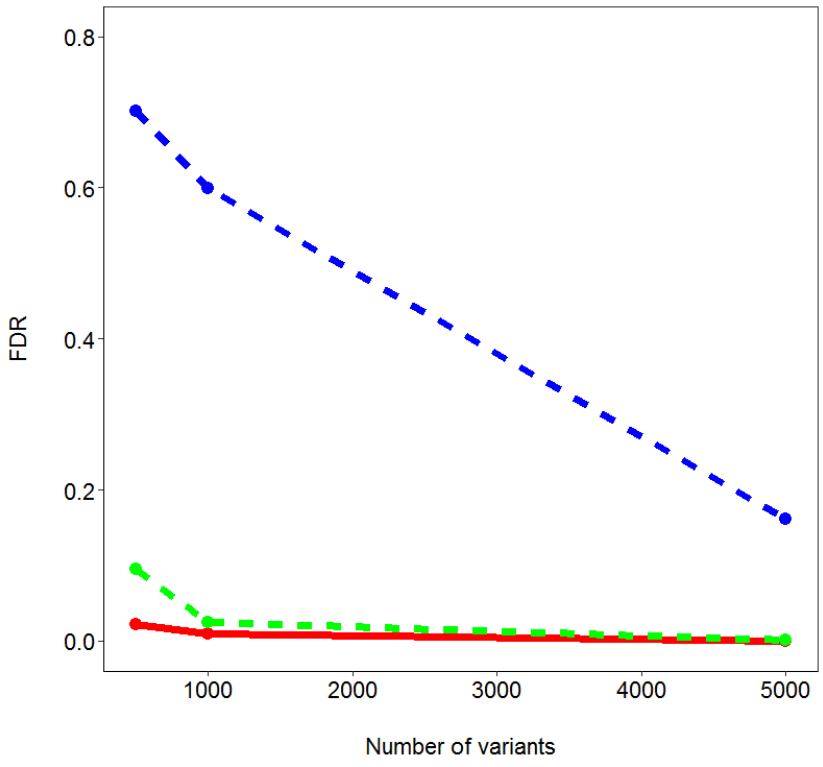


Figure 3.3: False discovery rate of our method (segment size =3) with different variant-selection criteria vs. KING without encryption. The green and blue dotted line represent results using our method with variants selected by AFs and variants randomly selected, respectively. The red solid line represents the results using KING without encryption.

3.3.4 Comparison of the encryption schemes with different segment sizes

In this analysis, we investigate the utility and security of the encryption schemes with different segment sizes. Previously, we evaluated the encryption scheme with a segment size of 3 (Table 3.6). In this section, the genotype is encrypted under another scheme with a segment size of 5. We select 500, 1000 and 5000 variants based on criterion-2. Table 3.7 summarizes the relative pairs inferred using this scheme.

Table 3.3 and Supplementary Table S3.1 show the data compression from the original genotype segments to the encryption codes for these two schemes. In terms of security, data is compressed more in the scheme with a segment size of 5 than 3 (segment size = 3: 27 values to 5 values vs.

segment size = 5: 243 values to 9 values). As a consequence, the scheme with segment size of 5 is more secure.

In terms of utility, the scheme with segment size of 3 performs better, especially with a smaller number of variants, as shown in Figure 3.4. The FDR is 9.51% for inferring 1st degree relatives using 500 variants for the scheme with a segment size of 3 while the FDR for the scheme with a segment size of 5 is 68.35%. As the number of variants increases, the issue of the inflated false discovery rate is resolved in 1st degree relative inference for the scheme with a segment size of 5 (FDR = 3.50%). However, the inference of 2nd degree relatives still suffers from the compressed information for this scheme. The FDR is 92.33% with 5000 variants (Supplementary Table S3.4). In other words, the scheme with a segment size of 5 does not have the ability to infer 2nd degree relatives with fewer than 5000 variants.

Table 3.7: Number of relative pairs inferred in a heterogeneous population using our method (segment size =5) with variants selected based on criterion-2. Data in the bracket represents number of correctly inferred pairs compared with gold-standard.

Number of variants	Number of pairs inferred (correctly inferred)
Duplicates/MZ twins (Gold standard =60)	
500	60(60)
1000	60(60)
5000	60(60)
1st degree relatives (Gold standard = 3425)	
500	10498(3323)
1000	3611(3413)
5000	3544(3420)

3.3.5 Two-step computational strategy and computational cost of the protocol

Recall the form of the log-likelihood,

$$l(\boldsymbol{X}_i, \boldsymbol{G}_j | R, \varepsilon) = \sum_p log(Pr(X_{ip}, \boldsymbol{G}_{jp} | R, \varepsilon)).$$

Here we take the scheme with segment size 3 as an example. The encrypted genotype code X_{ip} is element in {1, 2, 3, 4, 5, missing}. The genotype segment of 3 variants takes 4^3= 64 distinct values since each genotype in the segment is element in {0, 1, 2, missing}. As a consequence, the $(X_p, \boldsymbol{G}_p)$ pair takes a value in a limited set that has 384 distinct values. Given a relationship and segment p, the probability $Pr(X_p, \boldsymbol{G}_p | R, \varepsilon)$ for a certain $(X_p, \boldsymbol{G}_p)$ pair should be the same no matter which 2 samples we compare. In other words, when inferring relationships between large cohorts, the same $Pr(X_p, \boldsymbol{G}_p | R, \varepsilon)$ is used repeatedly.

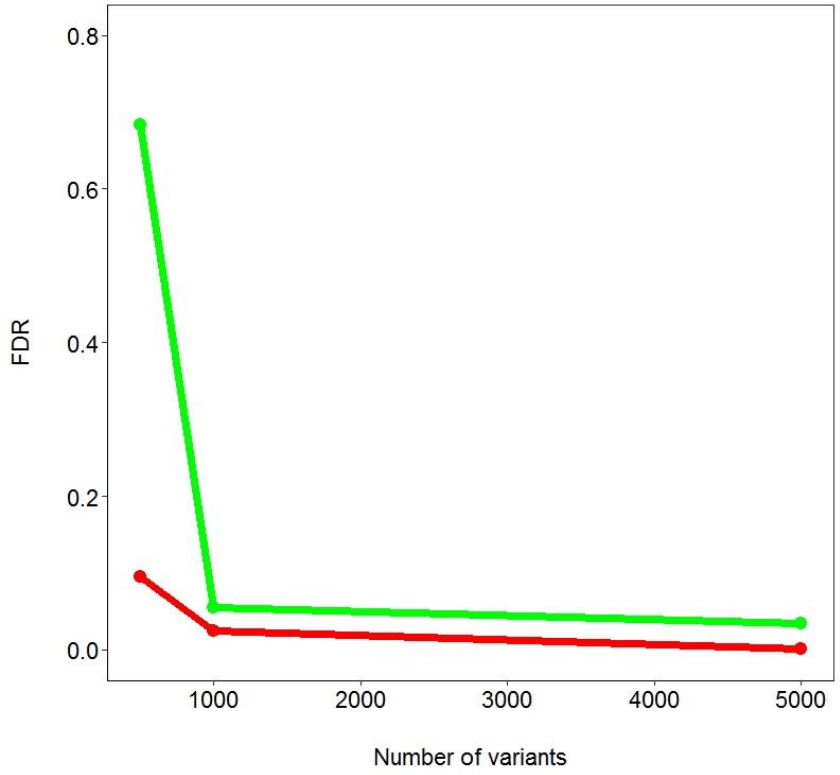


Figure 3.4: False discovery rate of our method with segment size = 3 vs. segment size = 5. The red and green line represent results using our method with segment size equals 3 and 5, respectively.

In order to make the protocol more efficient to deal with large cohorts, the calculation of the likelihood is conducted in two steps. In step one, we calculate the probability $Pr(X_p, \boldsymbol{G}_p | R, \varepsilon)$ for

all possible $(X_p, \boldsymbol{G}_p)$ pairs, all relationships and all segments. The results are saved with unique labels depending on p, R, and values of X and G. Then in step two, we calculate the likelihood for each pair of samples i and j, $\sum_p log(Pr(X_{ip}, \boldsymbol{G}_{jp}|R, \varepsilon))$, by direct calling the value of $Pr(X_p, \boldsymbol{G}_p|R, \varepsilon)$.

Computational time and memory usage is measured by applying the protocol on the heterogeneous population of 7,113 samples using variants selected based on criterion-2. Table 3.8 shows the computational time of each step in the protocol under different scenarios.

For study A, the encryption only takes several seconds for 10000 samples (0.8s for 500 variants and 10.5s for 5000 variants).

The step of the inference conducted by study B which contains two steps is more time-consuming. The computational time of step one does not depend on the number of samples we investigate. For scheme1, it scales well in practice, taking less than one minute for 500 variants and 1.9 CPU hours for 5000 variants. The computation time grows approximately quadratically with the number of variants. For step 2, the computation time of each pair of sample grows linearly with the number of variants while the total computation time grows linearly with sample sizes in Study A and B. For example, using 500 variants, each comparison takes 0.89ms. Then for two studies both having 10000 samples, the total comparison across studies requires 24.7 CPU hours. Compared with the computing time for step 2, the computational overhead arising from step 1 can be ignored. With 500 variants using scheme1, the whole process takes about 25 CPU hours.

For scheme 2, since both the encrypted genotype code $X_{ip} \in \{1,2,..,9, missing\}$ and the true genotype sequence (4^5= 1024) have more distinct values, step 1 is less efficient compared with

scheme1, taking 14 minutes for 500 variants and 26.5 CPU hours for 5000 variants. For step 2, it is also less efficient. Based on our analysis in Section 3.3.4, with scheme 2, we need 5000 variants to get a reasonable inference, which will take 8932.5 CPU hours for step 2, with the whole process taking 8959 CPU hours for 10000 samples vs. 10000 samples comparison.

In terms of memory usage, the pre-calculated values in step 1 need to be stored determined by the size of the segment and number of variants. For the encryption scheme with a segment size of 3, on average it requires 1.5MB memory to store the values for 500 variants and 14.6MB for 5000 variants. For the scheme with segment size of 5, it requires 234.4MB and 2343.8MB respectively.

Table 3.8: Computing time of each step in the protocol and memory usage in Step1 for different encryption schemes.

Scheme	Number of variants	Encryption (per sample)	Inference-Step1	Inference-Step2 (per pair)	Memory usage in Step1
Scheme1-size 3	500	0.08ms	52.68s	0.89ms	1.5MB
	1000	0.17ms	209.19s	1.92ms	2.9MB
	5000	1.05ms	6717.01s	11.90ms	14.6MB
Scheme2-size 5	500	0.07ms	850.86s	29.78ms	234.4MB
	1000	0.13ms	2938.55s	49.37ms	468.8MB
	5000	0.87ms	95427.43s	321.57ms	2343.8MB

3.4 Discussion

In this chapter, we have proposed a protocol to infer genetic relatives between studies without compromising privacy. The protocol depends on a likelihood-based model to infer relationships based on genotype data. It encrypts individual-level genotypes by dividing genotypes into segments and using summary statistics to represent the information in each segment.

Demonstrated in the real data simulation using TOPMed samples, our novel protocol is able to identify most of the 1st degree relatives while controlling the false discovery rate well with randomly selected variants in homogenous populations. By applying the two-step strategy for likelihood calculation, we show that the secure protocol scales well for large-scale studies. By comparing two encryption schemes with different segment sizes, we demonstrate that they both have pros and cons in terms of security, utility and computation efficiency.

The inference of relatedness has been shown to be highly associated with allele frequencies. With the development of genotyping technologies, quite a few methods have been proposed to infer genetic relationships based on allele frequencies of the genotypes. For example, one kind of method infers relationships by calculating relatedness coefficient (Lynch, 1988; Queller and Goodnight, 1989; Lynch and Ritland, 1999; Wang, 2002; Milligan, 2003; Thomas, 2010). Another kind of method infers relationships more accurately using a likelihood ratio model that compared multipoint probability of markers conditional on relationships (Boehnke and Cox, 1997; Broman and Weber, 1998; Epstein et al. 2000). The inference of relatedness based on all of these methods has been shown to be highly correlated with allele frequencies. The methods are only consistent or unbiased under the assumption that the allele frequencies are known without errors (Lynch and Ritland, 1999; Wang, 2002). Previous studies have shown that when biased reference allele frequencies are incorporated in practice, the estimates of relatedness are biased, leading to inaccurate relationship inference (Anderson and Weir, 2007; Wang, 2014; Wang, 2017). The problem of the inaccurate allele frequency, similarly, leads to the biggest limitation of our method about inferring relationships in the population with diverse ethnic backgrounds. One essential assumption of our method is that allele frequencies of the variants should be the same across all the samples in the study. This assumption is violated in

heterogeneous populations where allele frequencies vary across distinct ethnic groups. If applying the allele frequencies of the joint population, we will get a biased likelihood, which endangers our relationship inference. As a consequence, we are more likely to infer an unrelated pair to be relatives. To diminish the effect of population stratification, we introduce a variant selection strategy. We select variants that have consistent allele frequencies across different ethnic groups to infer the relationships. In this way, we resolve the problem to the extent that we can infer the 1st degree relatives with high accuracy and a controlled false discovery rate. However, this solution does have limitation in solving real data problems when retrieving the ethnic information requires a lot of effort. The variant selection strategy only solves the problem under certain scenarios.

Since more and more reliable reference recourses of human variants have become available in database like dbSNP (Sherry et al., 2001), we can also address this problem by applying ethnic-specific AFs reported in the reference recourses to our model. Such AFs may better represent the true AFs of each individual, so that the issue of the inflated false discovery rate will be resolved. In order to assign the correct ethnic-specific AFs to each individual, the ancestry information should be shared between studies. A caveat is that for an admixed sample, a method to calculate his/her AFs accurately based on the AFs within each ethnic group is required.

Other than assisting our inference in heterogeneous population, the reference allele frequency, on the other hand, may also lead to attacks to our encryption scheme. Homer et al. (2008) claimed that, under certain conditions, by comparing the MAFs of a specific individual to the distribution of MAFs in a reference population, they could use statistical methods to infer the presence of an individual with known genotype in a mix of DNA. It raises the concern about the security of sharing summary statistics between studies. In our protocol, we encrypt genotypes using the

summary statistics. Even though we demonstrate the underlying individual-level genotype cannot be disclosed directly from the encrypted code, our protocol still has the risk of information leakage when we compare the summary statistic against the reference AFs. For example, if a variant has low AF in the reference, some segments with genotype equals 0 for that variant may have higher probability than others. As a consequence, the genotype information may be disclosed when only one segment has genotype that equals 0 at that position among all segments corresponding to a certain encrypted code. Here, we protect our protocol against such attack by avoiding using variants with lower AFs. Even with the reference information, we cannot guess with confidence of the genotype for a common variant with AF around 0.5. Another attack occurs when variants within one segment are in high linkage disequilibrium (LD). In this situation, segments having the same genotype at each position, like 000, 111 and 222, have much higher probability than others. In our protocol, we address this issue by only examining independent variants. In the future, we may protect our scheme against these attacks using the differential privacy technique (Uhlerop et al., 2013; Yu et al., 2014). It adds reasonable noise to the summary statistic before its release. Then, the likelihood based on the genotype summary statistic with noise should be constructed to infer the relationship.

In conclusion, we propose a privacy preserving protocol that enables the relationship inference between studies without disclosing individual-level data. The method has limitations in inferring relationships in population with diverse ancestry when the ancestry information is not known. Thus, we need further investigation to make the protocol more practical for heterogeneous population. In the next chapter, we will propose another privacy preserving method that can infer relationships robustly in heterogeneous populations.

Supplements

Supplementary Table S3.1: Mapping between encrypted genotype and true genotype for encryption scheme with segment size = 5. Information is compressed from 243 distinct values into 9 distinct values.

Genotype segment (G)	Encrypted genotype code (*X*)
00000, 00001, 00010, 00100,01000,10000	1
00011, 00101, 01001,10001,00110,01010,10010,01100,10100,11000, 00002, 00020, 00200,02000,20000	2
12000, 20100, 11100, 02100, 10200, 01200, 20010, 21000, 11010, 02010, 10110, 01110, 00210, 10020, 01020, 00120, 20001, 11001, 02001, 10101, 01101, 00201, 10011, 01011, 00111, 00021, 10002, 01002, 00102, 00012	3
22000,21100,12100, 20200, 11200, 02200, 21010, 12010, 20110, 11110, 02110, 10210, 01210, 20020, 11020, 02020, 10120, 01120, 00220, 21001, 12001, 20101, 11101, 02101, 10201, 01201, 20011, 11011, 02011, 10111, 01111, 00211, 10021, 01021, 00121, 20002, 11002, 02002, 10102, 01102, 00202, 10012, 01012, 00112, 00022	4
22100, 21200, 12200, 22010, 21110, 12110, 20210, 11210, 02210, 21020, 12020, 20120, 11120, 02120, 10220, 01220, 22001, 21101, 12101, 20201, 11201, 02201, 21011, 12011, 20111, 11111, 02111, 10211, 01211, 20021, 11021, 02021, 10121, 01121, 00221, 21002, 12002, 20102, 11102, 02102, 10202, 01202, 20012, 11012, 02012, 10112, 01112, 00212, 10022, 01022, 00122	5
22200, 22110, 21210, 12210, 22020, 21120, 12120, 20220, 11220, 02220, 22101, 21201, 12201, 22011, 21111, 12111, 20211, 11211, 02211, 21021, 12021, 20121, 11121, 02121, 10221, 01221, 22002, 21102, 12102, 20202, 11202, 02202, 21012, 12012, 20112, 11112, 02112, 10212, 01212, 20022, 11022, 02022, 10122, 01122, 00222	6
22210, 22201, 22111, 21211, 12211, 22120, 21220, 12220, 22021, 21121, 12121, 20221, 11221, 02221, 22102, 21202, 12202, 22012, 21112, 12112, 20212, 11212, 02212, 21022, 12022, 20122, 11122, 02122, 10222, 01222	7
22220, 22211, 22121, 21221, 12221, 22202, 22112, 21212, 12212, 22022, 21122, 12122, 20222, 11222, 02222	8
12222, 21222, 22122, 22212,22221, 22222	9

Supplementary Table S3.2: Number of 2nd degree relative pairs inferred in a heterogeneous population using our method (segment size = 3) with randomly selected variants vs. using KING without encryption. Data in the bracket represents number of correctly inferred pairs compared with gold-standard. The gold standard is 3127 pairs.

Method	Number of variants	Number of pairs inferred (correctly inferred)
Our method w/ randomly selected variants	500	915298(2191)
	1000	315272(2586)
	5000	339785(2612)
	10000	216538(2689)
KING w/o encryption	500	38698(2355)
	1000	4316(2615)
	5000	3137(2978)
	10000	3142(3018)

Supplementary Table S3.3: Number of 2nd degree relative pairs inferred in a heterogeneous population using our method (segment size = 3) with variants selected based on different criteria. Data in the bracket represents number of correctly inferred pairs compared with gold-standard. The gold standard is 3127 pairs.

Method	Number of variants	Number of pairs inferred (correctly inferred)
Our method w/ criterion 1	500	257063(2430)
	800	54505(2584)
Our method w/ criterion 2	500	287799(2490)
	1000	27546(2810)
	5000	5014(3094)

Supplementary Table S3.4: Number of 2nd degree relative pairs inferred in a heterogeneous population using our method (segment size = 5) with variants selected based on criterion 2. Data in the bracket represents number of correctly inferred pairs compared with gold-standard. The gold standard is 3127 pairs.

Number of variants	Number of pairs inferred (correctly inferred)
500	1403761(2028)
1000	290618(2625)
5000	37485(2874)

CHAPTER IV

Robust Method for Identifying Genetic Relatives between Studies without Compromising Privacy

4.1 Introduction

As described in the previous chapter, personal genomics in the setting of electronic health records (EHR) studies has gained much interest recently due to the need to infer genetic relatives between studies. In Chapter III we proposed a secure protocol that can infer close relatives with high accuracy without disclosing individual-level data. However, its performance in dealing with a heterogeneous population is deteriorated by the differences in allele frequency among ancestry groups. In addition, none of the existing methods mentioned in Chapter III, including the private set intersection (De Cristofaro and Tsudik, 2012), oblivious transfer-based hamming distance system (Bringer et al., 2013), privacy-preserving approximating edit distance (Wang et al., 2015) as well as the "fuzzy" encryption methods, have been evaluated under multi-ethnic scenarios, i.e. heterogenous population. In this chapter, we try to overcome this limitation and propose a novel protocol that allows detection of genetic relatives among multi-ethnic groups without compromising privacy.

Our protocol securely infers genetic relatives by combining the robust relationship inference method previously described by Manichaikul et al. (2010) and an encryption technique called homomorphic encryption (Gentry, 2009; Fan and Vercauteren, 2012). It has several advantages.

First, our protocol only requires the sharing of a limited number of variants so that the computation of our protocol scales well in practice. Second, our protocol is robust to population stratification. The method proposed by Manichaikul et al. (2010) infers genetic relatives using kinship coefficients. It has been shown to have reliable performance even under scenarios where population stratification and violation of Hardy-Weinberg equilibrium (HWE) are present. Moreover, the security of our protocol is guaranteed theoretically by the rigorous proof for homomorphic encryption (Gentry, 2009).

Homomorphic encryption is a form of encryption that allows one to conduct calculations on encrypted data without first decrypting the data where only people with the decryption key can decrypt the result. The encrypted result, when decrypted, matches the result of calculations performed on the real data. Gentry first constructed the fully homomorphic encryption (FHE) scheme which enables one to perform arbitrary computations on encrypted data; however, implementations of FHE are generally inefficient (Gentry, 2009). More recently, a more practical scheme called the somewhat homomorphic encryption scheme (SWHE) was proposed by Fan and Vercauteren (2012). It allows us to evaluate a limited number of operations, which is proved to be sufficient for calculating kinship coefficient in our framework.

Previously, practical homomorphic encryption schemes have been widely applied to genetic data. Much of the work primarily focuses on the problem of pattern matching for genomic sequences. For example, Blanton et al. (2012) proposed a method for genome sequencing comparison using garbled circuits. Cheon et al. (2015) proposed a method to calculate edit distance on homomorphically encrypted data. Later, the efficiency of edit distance calculation and string searching were improved by using more efficient homomorphic encryption schemes (Kim and Lauter, 2015; Shimizu et al., 2016). On the other hand, homomorphic encryption has also been

applies to conduct statistical tests in genetics. One study presented the encryption scheme that allowed secure outsourcing of GWAS results to external data center (Kantarcioglu et al., 2008). Ayday et al. (2013) showed how to use additive homomorphic encryption to predict disease susceptibility securely using genetic information. In addition, Lauter et al. (2014) conducted genetic tests for HWE and linkage disequilibrium (LD) securely using the leveled homomorphic encryption scheme. Later, by incorporating an honest-but-curious key manager, Ugwuoke et al. (2017) improved the efficiency of the test for LD. The method proposed by Kim and Lauter (2015) showed the utility of homomorphic encryption in calculating minor allele frequency and chi-square statistic in GWAS. Homomorphic encryption has been applied to solve problems in multiple fields of genetic studies, however, for our particular purpose of inferring genetic relatives, the method based on this technique is still lacking.

Thus, in this Chapter, we address the problem of relationship inference between studies by applying homomorphic encryption in our protocol. Kinship coefficients can be obtained without disclosing any genetic information between studies. Through simulations, we show that our protocol successfully encrypts genetic data and decrypts kinship coefficient results under different scenarios. Our performance is identical to using KING with unencrypted genotypes. Furthermore, we demonstrate the utility of our method by applying our technique to infer genetic relatives among samples from the Trans-Omics for Precision Medicine (TOPMed) study and the Genome Aggregation Dataset (gnomAD) (Karczewski et al., 2019; Taliun et al., 2019).

4.2 Method

4.2.1 Robust genetic relative inference in the presence of population structure

First, we describe how to infer genetic relatives with unencrypted data in diverse ethnic groups, using a robust method proposed by Manichaikul et al. in 2010. It is implemented in the KING software. For simplicity, we call this method KING in this chapter.

Let ϕ_{ij} denote the kinship coefficient between sample i and j, which is the probability of two alleles randomly selected from two individuals are identical by descent. KING provides a robust estimator for ϕ_{ij} based on M pairs of variants without missing genotypes in both individual i and j. The details of deriving the robust estimator are provided in the Supplementary note 1.

The robust estimator of the kinship coefficient is

$$\widehat{\phi_{IJ}} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)} + N_{Aa}^{(j)}},$$

where $N_{Aa}^{(i)}$ is the total number of heterozygotes for the *i*-th individual among the M variants. $N_{Aa,Aa}$ is the total numbers of variants at which the individuals of the pair are heterozygous. Finallly, $N_{AA,aa}$ is the total numbers of variants at which individuals of the pair are homozygous of different alleles.

In the paper by Manichaikul et al, a more robust estimator was proposed to deal with a situation when the violation of HWE of some variants results in excessive heterozygosity.

When the assumption of HWE is violated, the robust estimator for ϕ_{ij} is

$$\widehat{\phi_{IJ}} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{2\min(N_{Aa}^{(i)}, N_{Aa}^{(j)})} + \frac{1}{2} - \frac{1}{4}\frac{N_{Aa}^{(i)} + N_{Aa}^{(j)}}{\min(N_{Aa}^{(i)}, N_{Aa}^{(j)})}.$$

Once we get the kinship coefficient, the relationship can be inferred based on the criteria in Table 4.1.

Table 4.1: Relationship inference criteria for kinship coefficient.

Relationship	ϕ Inference criteria
MZ twin	$>1/2^{3/2}$
Parent-offspring/Full-sib	$(1/2^{5/2}, 1/2^{3/2})$
Second degree	$(1/2^{7/2}, 1/2^{5/2})$
Unrelated	$<1/2^{9/2}$

4.2.2 Procedures of homomorphic encryption and a general secure relationship inference framework

To infer genetic relatives securely, we need to calculate the kinship coefficient defined above using encrypted genotype data. The homomorphic encryption technique enables us to perform the computation on encrypted data without knowing any decryption information. The general procedure of homomorphic encryption is shown in Figure 4.1. It includes:

Key-generation: generating public key (pk) and secure keys (sk) based on pre-specified parameters;

Encryption: encrypting plaintext to ciphertext using public key;

Evaluation: calculating on ciphertext, practical homomorphic encryption schemes only support addition and multiplication;

Decryption: decrypting ciphertext using secure key.

Based on these procedures, we propose a privacy-preserving framework that allows two studies to infer genetic relatives without exposing their individual-level genotype information. Two parties included in this protocol have such responsibilities:

Study A: generates the public and secure keys to encrypt data and decrypts the result.

Study B: calculates encrypted result using its own data and encrypted data from study A.

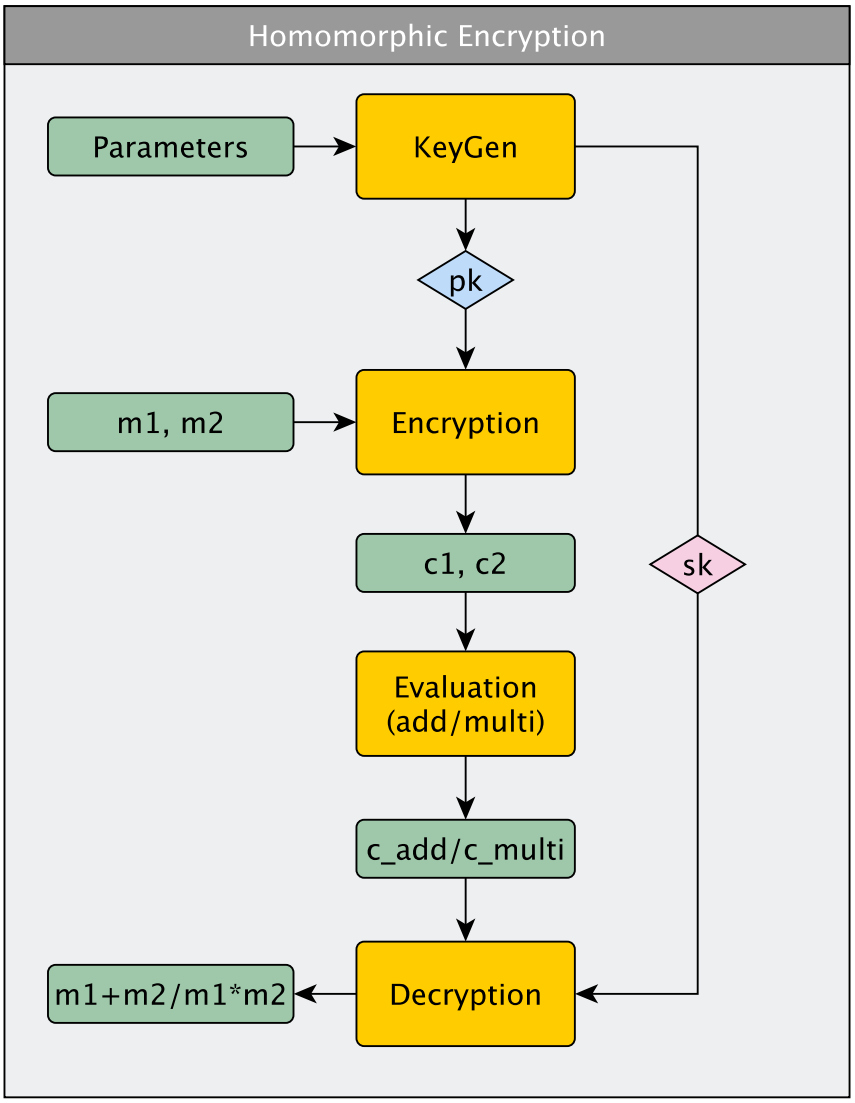


Figure 4.1: General procedures of homomorphic encryption. Here, pk and sk denote public key and security key, respectively; m1 and m2 denote the message before encryption; c1 and c2 denote the encrypted ciphertext; c_add and c_multi denote the encrypted results of addition or multiplication of the ciphertext.

The protocol follows these general steps:

1) Study A generates keys (pk,sk), encrypts its genotype using pk and sends it to study B.
2) Study B calculates encrypted kinship coefficients using encrypted genotype from study A and its own genotype. Then sends the encrypted result back to study A.

3) Study A decrypts the result to get kinship coefficient using sk and makes inference of the relationship.

4.2.3 Somewhat homomorphic encryption

Given that only a limited number of operations are needed for kinship coefficient calculation, we used the somewhat homomorphic encryption scheme proposed by Fan and Vercauteren (FV, 2012) for the relationship inference. This homomorphic encryption supports both addition and multiplication on the encrypted data. The details of this encryption scheme are shown below.

Notation and parameters

Our customized FV scheme operates in the ring $R \stackrel{\text{def}}{=} \mathbb{Z}[X]/(X^n + 1)$, whose elements are polynomials with integer coefficients of degree less than n. We call $X^n + 1$ the polynomial modulus. Usually n is set to be a power of 2. Messages (plaintext), encrypted messages (ciphertexts), public and secure keys are elements in the ring R. The notation [a]q is to denote the operation of reducing the coefficients of $a \in R$ modulo q into the set $\left(-\frac{q}{2}, \frac{q}{2}\right]$.

Suppose the plaintext space is $R_t \stackrel{\text{def}}{=} \mathbb{Z}_t[X]/(X^n + 1)$ whose elements are polynomials with integer coefficients modulo t. t is called the plaintext modulus. Suppose the ciphertext space is $R_q \stackrel{\text{def}}{=} \mathbb{Z}_q[X]/(X^n + 1)$ whose elements are polynomials with integer coefficients modulo q. q is called the coefficient modulus.

Let $\boldsymbol{x} \leftarrow D$ denotes that $\boldsymbol{x}$ is sampled from distribution D. Two distributions are relevant to our scheme, R_q and χ_{err}. Here R_q is a uniform distribution on the R_q space. For example, R_3 is the uniform distribution of polynomials with coefficients in {-1,0,1}. χ_{err} is the distribution of error that we add to encrypt the data. We use discrete Gaussian distribution with mean 0 and standard deviation σ for the distribution χ_{err}.

Encoding of the message

Since all operations are done in the ring space, to encrypt an integer μ, we should first encode it as a polynomial-plaintext $m \in R$. One general way is to take bit-decomposition of μ and use bits as coefficients of the polynomial. For instance, if we take base of 2, for integer $\mu = \sum_i \mu_i 2^i$ ($\mu_i \in \{0,1\}$), the corresponding plaintext should be $m = \sum_i \mu_i X^i$. However, in a later section we will show that the raw data used in our protocol are 0 and 1 which are already in the polynomial-plaintext format. The encoding step is not necessary for our protocol.

Formal definition of the steps in the encryption scheme

Then the FV encryption scheme is defined as follows:

KeyGen(params): sample $\boldsymbol{s} \leftarrow R_3$, $\boldsymbol{a} \leftarrow R_q$ and $\boldsymbol{e} \leftarrow \chi_{err}$. Then the public key (pk) and secure key (sk) should be

$$pk = \left([-\boldsymbol{a} \cdot \boldsymbol{s} + \boldsymbol{e}]_q, \boldsymbol{a}\right); \quad sk = \boldsymbol{s}.$$

Encrypt(pk,m): let m be the message and $\boldsymbol{m} \in R_t$, let $\boldsymbol{p_0} = pk[0], \boldsymbol{p_1} = pk[1]$, sample $\boldsymbol{u} \leftarrow R_3, \boldsymbol{e_1}, \boldsymbol{e_2} \leftarrow \chi_{err}$. Then message can be encrypted by

$$ct = \left(\left[\boldsymbol{p_0} \cdot \boldsymbol{u} + \boldsymbol{e_1} + \left\lfloor \frac{q}{t} \right\rfloor \cdot \boldsymbol{m}\right]_q, [\boldsymbol{p_1} \cdot \boldsymbol{u} + \boldsymbol{e_2}]_q\right).$$

Here by expanding the public key in the ciphertext, we get $ct[0] = \left[\boldsymbol{e} \cdot \boldsymbol{u} + \boldsymbol{e_1} - \boldsymbol{a} \cdot \boldsymbol{s} \cdot \boldsymbol{u} + \left\lfloor \frac{q}{t} \right\rfloor \cdot \boldsymbol{m}\right]_q$. The last term that contains the message is masked by the term $\boldsymbol{a} \cdot \boldsymbol{s} \cdot \boldsymbol{u}$ which has equivalently large polynomial degree as $\left\lfloor \frac{q}{t} \right\rfloor \cdot \boldsymbol{m}$. The second term of the ciphertext becomes $ct[1] = [\boldsymbol{a} \cdot \boldsymbol{u} + \boldsymbol{e_2}]_q$. Therefore if knowing the secure key $\boldsymbol{s}$, we can multiply $\boldsymbol{s}$ and $ct[1]$ and use it to remove the large term in $ct[0]$ to decrypt the message.

Decrypt(sk,m): let $\boldsymbol{c_0} = ct[0], \boldsymbol{c_1} = ct[1]$. Then decrypted ciphertext should be

$$[\left\lfloor \frac{t \cdot [\boldsymbol{c_0} + \boldsymbol{c_1} \cdot \boldsymbol{s}]_q}{q} \right\rceil]_t.$$

As shown in the encryption step, $[\boldsymbol{c_0} + \boldsymbol{c_1} \cdot \boldsymbol{s}]_q$ gives us $\left[\boldsymbol{e} \cdot \boldsymbol{u} + \boldsymbol{e_1} + \boldsymbol{e_2} \cdot \boldsymbol{s} + \left\lfloor \frac{q}{t} \right\rfloor \cdot \boldsymbol{m}\right]_q$, which only contains the message and random errors having coefficient much smaller than $\left\lfloor \frac{q}{t} \right\rfloor$. If we rescale the coefficients of this polynomial back to values in mod t and round them, we can remove the errors and recover the message m.

Add($\boldsymbol{ct_1}, \boldsymbol{ct_2}$): Given two ciphertext ct_1, ct_2, then the addition of them should be

$$ct_{add} \coloneqq \left(\left[ct_1[0] + ct_2[0]\right]_q, \left[ct_1[1] + ct_2[1]\right]_q\right).$$

Since later we show that our protocol does not depend on multiplication of the ciphertexts, details of the multiplication step are not provided here.

For this protocol, the FV encryption scheme is implemented using a C++ library, SEAL v2.3.0 (Bajard et al., 2016; Chen et al., 2017).

Selection of parameters

The utility, security and efficiency of the FV encryption scheme depend on the choice of encryption parameters.

First of all, each ciphertext contains noise, which grows in all homomorphic operations, and eventually reaches a maximum value. Once this maximum is reached, the ciphertext cannot be decrypted correctly. Thus to achieve the utility of the encryption scheme, the choice of parameters should guarantee that the maximum noise boundary is large enough for our predetermined operations.

In terms of efficiency, the operations on polynomials with smaller degree and smaller coefficients are more efficient. Given a certain level of security, it is important to set the parameters that can achieve the balance between efficiency and utility.

As demonstrated above, the encryption scheme contains four main parameters, polynomial modulus $X^n + 1$, plaintext modulus t, coefficient modulus q and standard deviation σ of the error distribution.

The polynomial modulus $X^n + 1$ mainly affects the security level of the scheme. The larger the n is, the more secure the scheme will be. In addition, larger n leads to larger ciphertext sizes and consequently slower operations. The coefficient modulus q affects the utility of the scheme. Larger q allows more complicated computations. However, a larger q also lowers the security level of the scheme. The plaintext modulus t has an opposite effect on the utility. Smaller t allows more complicated computations.

SEAL provides the default value of parameters for different security levels based on security level estimates (Chase et al., 2017). Later in our simulation study, we will provide the optimal parameter selection for our protocol under different situations.

4.2.4 Encryption of genetic data and secure relationship inference protocol

Since only addition and multiplication are supported by the FV encryption scheme, in order to calculate kinship coefficient, $\widehat{\phi_{ij}} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)} + N_{Aa}^{(j)}}$, we have to calculate the denominator and numerator separately.

To calculate the numerator, rather than encrypting directly on genotype $\{0,1,2\}$, we compute 3 ciphertexts for one variant which indicate the genotype equals 0, 1 and 2 respectively. In this way, the count numbers, $N_{Aa,Aa}$ and $N_{AA,aa}$ can be calculated by summing the corresponding

indicator products. To be specific, suppose genotypes are encrypted for sample i in study A. Let $c_0^{(i,m)}, c_1^{(i,m)}, c_2^{(i,m)}$ denote the ciphertext for individual i at variant m. Let $g^{(i,m)}$ denote the true genotype of sample i at variant m. And $I(g^{(i,m)} = k)$ is the indicator function of the genotype. Then three encrypted values for one variant are encoded as follows:

$$c_0^{(i,m)} = Encrypt(pk, I(g^{(i,m)} = 0)),$$

$$c_1^{(i,m)} = Encrypt(pk, I(g^{(i,m)} = 1)),$$

$$c_2^{(i,m)} = Encrypt(pk, I(g^{(i,m)} = 2)).$$

Let $g_0^{(j,m)}, g_1^{(j,m)}, g_2^{(j,m)}$ denote the true genotype indicator for study B. Then $N_{Aa,Aa} - 2N_{AA,aa}$ for sample i in study A and j in study B can be calculated as follows:

$$\widehat{N_{Aa,Aa} - 2N_{AA,aa}}^{(i,j)} = \sum_m c_1^{(i,m)} g_1^{(j,m)} - 2 * (\sum_m c_0^{(i,m)} g_2^{(j,m)} + \sum_m c_2^{(i,m)} g_0^{(j,m)}). \quad (1)$$

Since for study B, $g^{(j,m)}$ is a known value without encryption, this calculation only requires addition of the ciphertexts.

For the denominator, we treat $N_{Aa}^{(j)}$ as a sharable summary statistic. Study B should send the unencrypted heterozygous count $N_{Aa}^{(j)}$ to study A. Therefore study A can calculate $\frac{1}{N_{Aa}^{(i)} + N_{Aa}^{(j)}}$ and multiplied it by the statistic (1) to get the kinship coefficient.

In practice, individuals may have missing genotypes. In the definition of the kinship coefficient, all the calculations are done with M pairs of variants without missing genotypes in both individuals of a pair. The missing value does not affect the calculation of the numerator that the missing genotype results in three indicator values equals 0 so that variants having missing value in either individual are not included in the calculation of the numerator automatically. However,

missing information is needed to calculate $N_{Aa}^{(i)}$ and $N_{Aa}^{(j)}$. We assume the missing information is not as sensitive as the genotype so it can be shared between studies. Along with the encrypted and decrypted information, two studies should share vectors M_i and M_j indicating whether variants are missing in their data.

Then, the secure protocol of inferring relationships between two studies follows the following steps (illustrated in Figure 4.2):

1) Two studies share their missing data information for each variant.
2) Study A generates keys (pk,sk), encrypts their genotype to be C_0, C_1, C_2 using pk and calculates heterozygous count $N_{Aa}^{(i)}$ for each sample based on the missing information. Then study A sends the encrypted genotype to study B.
3) Study B uses their genotype G_0, G_1, G_2 and encrypted genotype from study A to calculate encrypted numerator of the kinship coefficient. Then calculates $N_{Aa}^{(j)}$ and sends the results to study A.
4) Study A decrypts the numerator using sk and calculates the kinship coefficient.

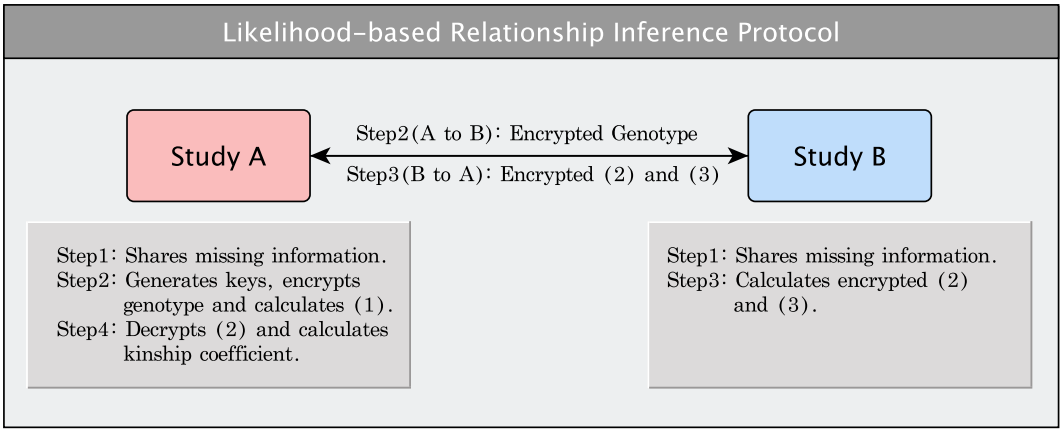


Figure 4.2: The specific process for securely calculating kinship coefficient between studies. Here, (1) denotes $N_{Aa}^{(i)}$; (2) denotes $N_{Aa,Aa}\widehat{-2}N_{AA,aa}^{(i,j)}$; (3) denotes $N_{Aa}^{(j)}$.

4.2.5 Security

The security of data of study A is always guaranteed by the security proof of homomorphic encryption (Gentry, 2009; Martin et al, 2015). Without knowing the security key, for study B, the encrypted genotype of study A is no more than a random value. However, data of study B may have the risk of getting disclosed when study A is an untrusted party. Here we demonstrate three possible attacks through which study A can disclose genotype of study B and we show how to protect data of study B again those attacks. To simplify the notation in this section, we combine three indicator vectors, g_0, g_1, g_2 for one sample i into one vector of length 3M, where M is the number of variants.

Attack I: Function privacy

The homomorphic encryption scheme implemented in SEAL makes no attempt to keep information secure from the owner of the secret key (Chen et al., 2017). In other words, the owner of the secret key, study A, can distinguish the noise from the output ciphertext, and deduce information about the function study B uses to get the output. For example, the highest power that is computed can be read from the size of the output ciphertext which may reveal the operations study B conducts.

To protect data of study B from disclosing, we incorporate a modified noise flooding method proposed by Gentry (Gentry, 2009). The main idea of this method is to mask the noise in the encrypted result that may reveal information of study B by adding encrypted zeros from B side. The encrypted zero should have noise that is polynomially equal or larger than the noise in the original encrypted result so that the old noise is flooded. To be specific, in our protocol, the noise in the encrypted $N_{Aa,Aa}$ and $N_{AA,aa}$ is result from 5M or less additions of the ciphertexts. Therefore in order to mask the noise in the true encrypted result, we added 5M encrypted zeros

to the encrypted result. The new encrypted result is $Enc\left(N_{Aa,Aa} - 2N_{AA,aa}\right) + \sum_{m=1}^{5M} Enc_m(0)$. Study A is still able to get the true value after decryption. At the same time, the noise in the encrypted zeros has similar order of magnitude as the noise in the true encrypted result. As a consequence, even with the security key, study A is not able to distinguish the noise of the true result from the noise in the zero-added result. Thus they are not able to uncover the data or function study B uses.

Attack II: Artificial genotype

In this protocol, two studies should share the summary statistic $N_{Aa,Aa} - 2N_{AA,aa}$ in order to calculate kinship coefficient. The summary statistic will not reveal specific genotype at each position if it is calculated based on the true indicator function of the genotype.

However, by encrypting 'artificial' genotypes instead of the indicator 0/1, study A may uncover the information of study B through the summary statistics. For example, study A encrypts $(1, a\ \ , \ldots, a^{M-1}, -0.5a^{M}, \ldots, -0.5a^{3M-1})$. Then the summary statistic, $N_{Aa,Aa} - 2N_{AA,aa}$, becomes $\sum_{m=0}^{M-1} {g_1}^{(j,m+1)} \cdot a^m + \sum_{m=M}^{2M-1} {g_0}^{(j,m+1)} \cdot a^m + \sum_{m=2M}^{3M-1} {g_2}^{(j,m+1)} \cdot a^m$. Getting the true genotype of study B from this summary statistic is just converting an integer from base-10 to base-a. Since study B cannot tell whether study A is cheating from the encrypted data without the secure key, we propose a modified protocol guaranteeing that study A can get the true value of the summary statistic only when it does not cheat.

Suppose $N_{Aa,Aa} - 2N_{AA,aa}$ is calculated based on M variants, by definition, this summary statistic should be an integer in [-2M, M], which contains 3M+1 different integers in total. So that if we add a random multiple of the integer 3M+1 to the $N_{Aa,Aa} - 2N_{AA,aa}$, then take modulus 3M+1 and map the value back to [-2M, M], we should be able to get the true value of the

summary statistic. However, if study A uses the 'artificial' information, $(1, a \ , \ldots, a^{\mathrm{M}-1}, -0.5a^{\mathrm{M}}, \ldots, -0.5a^{3\mathrm{M}-1})$, the value of the summary statistic will fall outside the range [-2M,M] that cannot be recovered by taking modulus 3M+1. In other words, the information will not be disclosed to study A if it uses the 'artificial' information. The modified steps include:

1) In step3, study B generates a random integer r from a discrete uniform distribution U(0,3M) and adds encrypted r*(3M+1) to the encrypted $N_{Aa,Aa} - 2N_{AA,aa}$.

2) In step4, study A decrypts the result and calculates $N_{Aa,Aa} - 2N_{AA,aa}$ + r*(3M+1) mod (3M+1) to get the true value of $N_{Aa,Aa} - 2N_{AA,aa}$.

Attack III: Aggregating information of multiple queries

Suppose study A keeps sending queries to get comparisons between its query sample and a sample from study B, once it gets enough information for a certain sample j in study B, study A is able to disclose the genotype of sample j by solving a linear system (Figure 4.3).

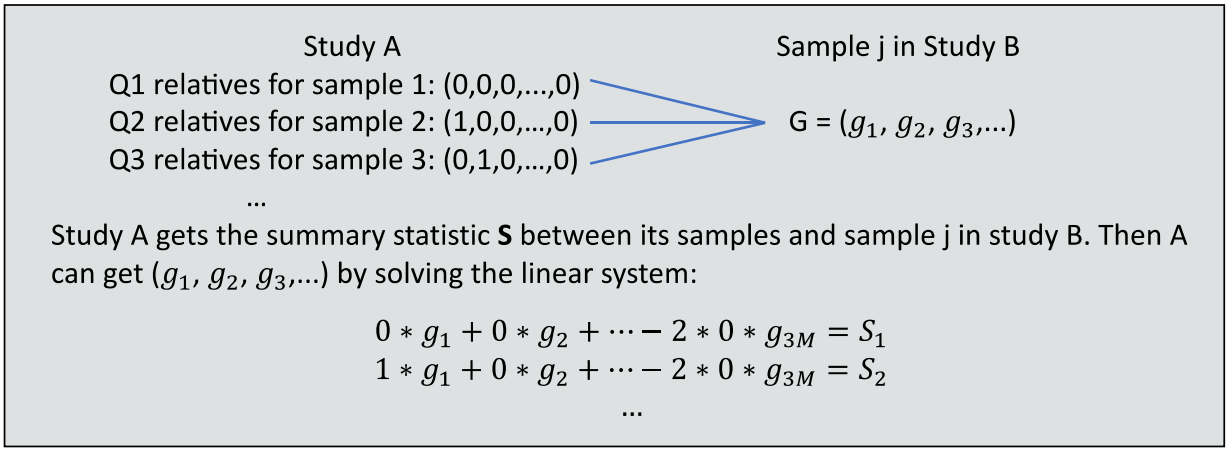


Figure 4.3: Attack on data of study B when study A gets enough information for a certain sample in study B.

To guard our protocol against this attack, after step 2, study B should generate a random order of its samples and send permuted results for each query. Study A is blind to the permuted order. In other words, study A cannot match the result of each query for a particular sample in study B. Figure 4.4 illustrates the permutation process. In the modified scheme, study B protects its genotype from being disclosed while study A can still infer whether its sample has relatives in study B.

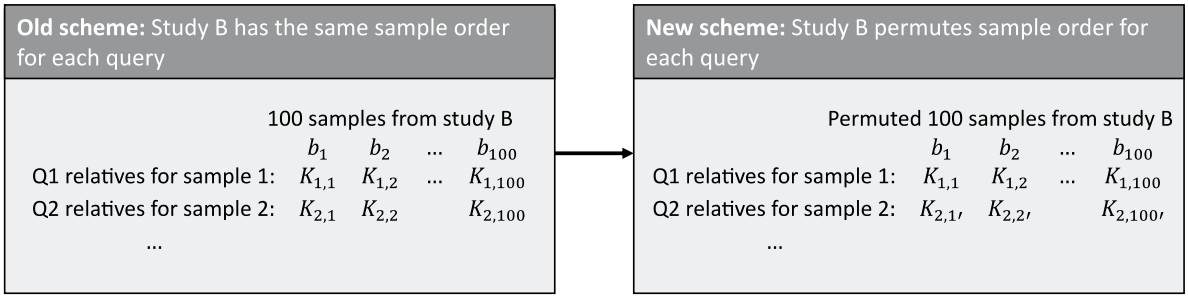


Figure 4.4: Permutation step when sending encrypted results back. $k_{i,j}$ represents the kinship coefficient for sample i in study A and sample j in study B.

One concern is the missing information shared between studies may potentially reveal some identification information of samples of study B. If the missing pattern is quite unique to each individual in study B, study A can use it to arrange the permutated result back to the original order.

In order to tackle this problem, study B should not share their missing information with study A; and study A should calculate an approximation of $N_{Aa}^{(i)}$ assuming no missing value in study B. Discussion about how ignoring the missing value in study B affects the genetic inference is shown in the Supplementary note 2. Through simulations we demonstrate that ignoring the missing information of study B will have negligible impact on the relationship inference

accuracy. Thus to prevent study A from identifying specific individual of study B in each query, we suggest study B not to share its missing information with study A.

4.3 Results

We evaluate the performance of this secure relationship inference protocol through several simulations. Again, relationship is inferred among the 14,572 samples of NHLBI TOPMed program Freeze3 data that we used in Chapter III. As introduced previously, BRAVO is the reference datasets of human genetic variation constructed based on TOPMed program (Taliun et al., 2019). Later in the data application, we will show how we apply the protocol to help with aggregating information of two similar reference datasets, BRAVO and gnomAD.

4.3.1 Performance of identification of the genetic relatives in homogeneous populations

The first simulation evaluates the protocol of inferring relationships when samples are having the same ethnic background. It assesses the number of variants that are needed to infer certain genetic relatives for our purpose of finding the overlapping samples between studies. The individual we use are exactly the same as Section 3.3.2, 3,357 Europeans, 3,437 Africans, 265 Asian and 54 Native Americans. Then relationships are inferred within each ethnic group. Gold-standard relationships among samples are inferred using KING based on ~600,000 HGDP variants which are considered to have high genotype quality (Cavalli-Sforza , 2005; Manichaikul et al. 2010).

We select different numbers of variants to infer relationships using our secure protocol and compare the performance with the gold-standard. For each pair of samples, we assume the sample with smaller ID is from study A; it follows the protocol by encrypting its genotype of the

selected variants and decrypting the result to infer the relationship. The sample with larger ID is from study B; it calculates the encrypted result using encrypted data from study A and its own genotype.

Analysis is conducted on independent 500, 1000, 5000 and 10000 variants with minor allele frequencies (MAFs) between 0.4 to 0.5. Again, to achieve our goal of identifying overlaps between the two studies, our primary concern is detecting duplicates and 1st degree relatives.

First of all, we compare the kinship coefficient calculated under our secure protocol with those calculated without encryption. The kinship coefficient obtained through the encryption/decryption process is identical as those calculated using the same genotypes without encryption. We demonstrate that through this protocol we successfully encrypt genotype data then successfully decrypt the results.

Table 4.2 shows the number of relative pairs inferred totally and inferred correctly with different numbers of variants for each group. All of the duplicate pairs are correctly inferred using 500 or more variants. For 1st degree relatives, we do not get false positive results in the Asian and Native American population. For the European and African population, the number of correctly identified relative pairs and the false discovery rate are comparable. For example, using 500 variants, we are able to identify 99.08% of the 1st degree relatives and get 2.56% false-positive pairs for Europeans while identify 99.32% 1st degree relatives with 1.63% false-positive pairs for Africans. And we get ideal results that identify all the true 1st degree relatives without any false positives using more than 5000 variants.

When using more than 5000 variants, 2nd degree relatives can also be identified with high discovery rate and low false discovery rate. Using 10000 variants, we detect 95.66% of the 2nd

degree relatives in Europeans with 5.07% FDR. In addition, we detect 98.19% of the 2nd degree relatives in Africans with 1.72% FDR.

Table 4.2: Number of relative pairs inferred with 500, 1000, 5000 and 10000 variants within each ethnic group. Data in the bracket represents number of correctly inferred pairs compared with gold-standard.

	European(N=3357)	**African(N=3437)**	**Asian(N=265)**	**American(N=54)**
Duplicates/MZ twins				
Gold standard	30	30	0	0
500	30(30)	30(30)	0	0
1000	30(30)	30(30)	0	0
5000	30(30)	30(30)	0	0
10000	30(30)	30(30)	0	0
1st degree relatives				
Gold standard	1959	1461	0	0
500	1992(1941)	1475(1451)	0	0
1000	1972(1951)	1461(1456)	0	0
5000	1959(1959)	1462(1461)	0	0
10000	1959(1959)	1461(1461)	0	0
2nd degree relatives				
Gold standard	2074	1049	1	0
500	17904(1549)	20306(889)	195(0)	2(0)
1000	2714(1723)	1638(807)	6(1)	0
5000	2088(1954)	1045(1020)	1(1)	0
10000	2090(1984)	1048(1030)	1(1)	0

4.3.2 Performance of identification of the genetic relatives in a heterogeneous population

The second simulation evaluates the protocol of inferring relationships securely in heterogeneous population. The simulation is conducted on all the 14,572 TOPMed samples with diverse ethnic background. In addition to those 7,113 samples with determined ancestry, 7,459 more samples are from other or admixed populations. Here, we assume the ancestry information is not known when inferring the relatives. Other simulation settings and procedures are the same as the previous simulation in homogeneous population.

Table 4.3 shows the number of relative pairs inferred in the heterogeneous population with different numbers of variants. All of the duplicate pairs are correctly inferred using 500 or more variants. For 1st degree relatives, the number of correctly identified relative pairs increases and the false discovery rate decreases as we use more variants. We are able to identify 99.24% of the 1st degree relatives while getting 2.07% false-positive pairs using 500 variants. Using 10000 variants, we are able to recover almost all of the 1st degree relatives while control the FDR as low as 0.04%. When using more than 5000 variants, 2nd degree relatives can also be identified. With 5000 variants, 95.70% of the 2nd degree relatives are detected with 4.35% FDR. With 10000 variants, 97.35% of the 2nd degree relatives are detected with 2.86% FDR.

Table 4.3: Number of relative pairs inferred with 500, 1000, 5000 and 10000 variants in a heterogeneous population.

Number of variants	Number of pairs	Number of overlapping results with Gold-standard
	Duplicates	
Gold-standard	108	108
500	108	108
1000	108	108
5000	108	108
10000	108	108
	1st degree relatives	
Gold-standard	8284	8284
500	8395	8221
1000	8315	8258
5000	8285	8279
10000	8286	8283
	2nd degree relatives	
Gold-standard	7444	7444
500	142646	5667
1000	11282	6320
5000	7448	7124
10000	7460	7247

4.3.3 Selection of parameters for the encryption scheme

As described before, the selection of parameters for the encryption scheme determines the utility, the security as well as the efficiency of the protocol. We have to define mainly 3 parameters for the SWHE scheme; polynomial modulus $X^n + 1$, plaintext modulus t and coefficient modulus q. A standard deviation σ of the error distribution is also a parameter and it is fixed to be 3.19 in SEAL. In this simulation, we provide guidance about how to set the optimal parameters for the encryption scheme to guarantee the security and utility of this protocol and also make the protocol scale well in practice.

The security is determined by both the polynomial modulus $X^n + 1$ and the coefficient modulus q. The larger the n is, the more secure the scheme will be. On the contrary, the smaller the q is, the more secure the scheme will be. In SEAL, coefficient modulus q is composed of a product of multiple small primes $q_1 \times \ldots \times q_k$. SEAL provides the default value of parameters paris (n, q) for different security levels based on security level estimates (Chase et al., 2017). For 128-bits of security, if n is set to be 1024, q should be any product of those small primes but at most 29 bits long. If n is 2048, then q can be larger values of at most 56 bits long. For higher security level 192-bits, smaller q should be used which is at most 39 bits when n is 2048.

In terms of the utility of the scheme, we have to set proper q and plaintext modulus t to control the noise growth in ciphertext so that it can be decrypted successfully. Here t is set to be a power of 2. Bigger q and smaller t allows more complicated computations. In our protocol, inferring relationships with more variants requires a higher level of utility. As described previously, we have an upper boundary for q for certain security and n. To satisfy the utility demand, we have to set t as small as possible. However, t cannot be any small value. It has a lower bound which determines the range of the plaintext value. When calculating on more variants, the range of the

value of the statistic is larger that requires larger t. Therefore sometimes we have to increase q and n accordingly in order to meet the security and utility requirement.

Larger n results in larger ciphertext sizes and slower operations. Even though there is no harm of using larger n in terms of security and utility, smaller n is always preferred in terms of achieving proficiency of the scheme. In Table 4.4, we show the optimal sets of parameters for different number of variants. Based on the simulation, they are the most efficient combination of parameters while also guarantee the security and utility of the encryption scheme.

Table 4.4: Optimal set of parameters for homomorphic encryption under different security levels.

Number of variants	128-bits			192-bits		
	n	q(bit)	t	n	q(bit)	t
500	1024	29	2^{10}~2^{12}	2048	39	2^{10}~2^{22}
1000	2048	56	2^{12}~2^{30}	2048	39	2^{12}~2^{20}
5000	2048	56	2^{14}~2^{30}	2048	39	2^{14}~2^{18}
10000	2048	56	2^{16}~2^{30}	2048	39	2^{16}~2^{18}

4.3.4 Computational cost and bandwidth consumption of the protocol

Table 4.5 and Supplementary Table S4.3 show the computing time of each primary step in the protocol under different scenarios. Results in Supplementary Table S4.3, demonstrate that the computing time of the encryption of an individual number is mainly determined by the degree of polynomial modulo n (for example, n=1024: 1.26ms vs. n=2048: 2.45ms). In addition, the computational time of the decryption of a number is also determined by n (for example, n=1024: 0.12ms vs. n=2048: 0.24ms). Overall, the computational time of encryption and decryption scales relatively well in practice. If we have 10000 samples in study A and study B, for 500 variants at 128-bits security level, the encryption and decryption take 7.0 CPU hours and 3.3 CPU hours respectively (Table 4.5).

The main cost is the evaluation on the encrypted data. The evaluation conducted by study B includes two steps. The first step is to generate a pool of encrypted zeros. For our protocol, we set the size of the pool to be 25M. The first step of generating a pool of encrypted zeros is done once for all the comparisons. The second step is to calculate $Enc\left(N_{Aa,Aa} - 2N_{AA,aa}\right) + \sum_{m=1}^{5M} Enc_m(0)$. For each pair of comparisons, 5M encrypted zeros are randomly selected from the pool and added to the result to avoid potential information leakage of study B (illustrated in Section 4.2.5). Overall, with the function protection step, the computational time of evaluation scales well for 10000 variants. When we infer relationships among large cohorts, the evaluation overhead of step one can be neglected compared with the total evaluation time. For the analysis between two studies with 10000 samples each, the CPU hours for evaluation are 292.6 hours for 500 variants and 3232.2 hours for 10000 variants.

Table 4.5: Computational time of primary steps in the protocol with 128-bits of security.

Number of variants	Procedure	Computational time	CPU hours for 10000 vs. 10000 comparison
500	Encryption	1.26ms/entry	7.0h
	Evaluation overhead*	16.08s	16.08s
	Evaluation	10.53ms/comparison	292.6h
	Decryption	0.12ms/comparison	3.3h
1000	Encryption	2.51ms/entry	27.9h
	Evaluation overhead*	63.57s	63.57s
	Evaluation	21.34ms/comparison	592.8h
	Decryption	0.26ms/comparison	7.2h
5000	Encryption	2.55ms/entry	141.7h
	Evaluation overhead*	320.78s	320.78s
	Evaluation	53.89ms/comparison	1497.0h
	Decryption	0.26ms/comparison	7.2h
10000	Encryption	2.52ms/entry	280.0h
	Evaluation overhead*	633.49s	633.49s
	Evaluation	116.63ms/comparison	3232.2h
	Decryption	0.18ms/comparison	5.0h

* This step is to encrypt 25M zeros before the evaluation. Then 5M encrypted zeros are randomly selected to be added to the evaluation for function security. This step is only done one time by study B for all comparisons.

In terms of bandwidth consumption, the size of the encrypted data is determined by polynomial modulo parameter n. On average, each ciphertext requires 17kb to store when n equals 1024 and 33kb to store when n equals 2048. This causes both studies to be burdened with heavy communication overhead. This overhead limits the number of samples we can compare within each communication.

4.3.5 Application to combine TOPMed and gnomAD reference datasets

As mentioned above, TOPMed program aims to provide reference resources of human genetic variation through a web browser, BRAVO. On BRAVO, variants summaries like allele frequencies (AFs) and quality metrics are shared to help researchers interpret the function of disease-causing variants. The current BRAVO browser is built upon 62,748 freeze5 TOPMed samples which have diverse backgrounds.

The Genome Aggregation Database (gnomAD) is also a resource that provides similar summary information of variants as TOPMed (Lek et al., 2016). It aggregates and harmonizes both exome and genome sequencing data from a variety of large-scale sequencing projects. The data we use in this application are data for constructing its web browser which contains 123,136 exome sequenced individuals and 15,496 whole-genome sequenced individuals (Lek et al., 2016; Karczewski et al., 2019).

Since BRAVO and gnomAD are designed for similar purposes and have similar content, researchers from both programs consent to aggregate information from both sides to construct a federated database. Since many studies participate in both programs, the overlapping samples may bias the summary statistics if we combine the two datasets directly. Thus in this application we want to infer overlapping samples between these two programs which is the first step of their federation.

Relationships are inferred through the secure protocol we propose using 500 common variants at 128-bits security level. The parameters we use for the encryption scheme are n = 1024, q = 29 bits long and $t = 2^{10}$. To avoid the extreme heavy communication overhead of the comparison between 200,000 samples, we get the raw genotype of selected 500 common variants from both studies. Then the encryption-decryption process is conducted on one side. Overall, the encryption, evaluation and decryption procedures take 33 CPU hours, 25k CPU hours and 290 CPU hours respectively.

Overall, 5568 pairs of duplicates and 2269 pairs of 1st degree relatives are detected between TOPMed and gnomAD. The overlapping samples are mainly from the studies that participate in both programs, including 2487 samples from Atrial Fibrillation Genetics Consortium, 1725 samples from Framingham Heart Study and 1868 samples from Jackson Heart Study.

4.4 Discussion

Here, we propose a privacy preserving method in the context of KING and homomorphic encryption. This protocol allows us to address the inherent tension of data sharing privacy in personal genomics. Under this protocol, we are able to infer genetic relatives without exposing individual-level genetic data. The results are robust in the population with diverse ethnic background. In addition to the general data-sharing framework of homomorphic encryption, we modify the protocol according to the calculation of kinship coefficients and provide protection against several major attacks.

The development of practical homomorphic encryption schemes provides us a chance to establish a two-party secure protocol for conducting genetic tests. While previous studies have shown its utility in conducting statistical tests for HWE, LD and genetic-disease associations in a

GWAS setting, the method to infer relatives based on this technique has not been well established. Even though methods that calculate edit distance or perform string search are able to find similar patients between studies, they are not scalable for our purpose of finding relatives between large-scale studies, since the problems they solve are much more complicated and have deeper circuits than the problem we consider here. To the best of our knowledge, the efficient method of calculation of the edit distance takes 15 to 100 seconds for a pair of 5000 variants (Kim and Lauter, 2015); the string search of a sequence of 25 variants in 2000 genomes takes 10 seconds (Shimizu et al., 2016). None of these is applicable to our scenarios where both studies have more than 10000 samples. In our protocol, instead of comparing DNA sequences, we use a simpler summary statistic, kinship coefficient, to infer relatives. Thus, our protocol requires much less computing time for each comparison and scales well for large-scale studies. In addition, these existing methods can only infer duplicates while our method can infer relatives up to 2nd degree. Moreover, these methods only protect data from one side, while we provide several layers of protection of the data on both sides.

Another major asset of our protocol is the reliable performance in heterogeneous populations. Compared to the existing methods and the method proposed in Chapter III, our method can infer relationships robustly in a heterogeneous population without disclosing ancestry information between studies. Through simulations and an application on data from TOPMed and gnomAD program in a heterogeneous population, we demonstrate that we are able to recover most of the 1st degree relatives while controlling the false discovery rate well using a limited number of variants. The whole protocol scales well in practice with as many as 10000 variants.

However, the protocol does suffer from heavy communication overhead. The large size of the ciphertext has been a general problem of all the practical methods based on homomorphic

encryption. To guarantee the security of the encryption scheme, both the degree of the polynomial and the coefficients are quite large and random. An efficient way to store the ciphertext may be difficult to develop. Some studies have demonstrated that a hybrid encryption scheme of homomorphic encryption and other encryption technique, for example Advanced Encryption Standard, may solve this common practical problem in homomorphic encryption (Naehrig et al. 2011; Olumide et al. 2015; Alkady et al. 2018).

In conclusion, we propose a secure protocol that enables the relationship inference between studies without sharing individual level data. In the next chapter, we will discuss the limitations regarding the communication overhead that need further investigation to make the protocol more practical for large-scale studies.

Supplements

Supplementary note 1: Robust relationship inference method previously described by Manichaikul et al (2010).

Let ϕ_{ij} denote the kinship coefficient between sample i and j, which is the probability of two alleles randomly selected from two individuals to be identical by descent. Let π_{0ij}, π_{1ij}, π_{2ij} denote the probability that two individuals share 0,1 and 2 alleles identical by descent (IBD). They have the relationship $2\phi_{ij} = \frac{\pi_{1ij}}{2} + \pi_{2ij}$. Supplementary Table S4.1 lists the probabilities for genotype pairs of bi-allelic variants given their IBD status. Then the marginal probability of genotype pairs can be represented by kinship coefficient:

$$\Pr(Aa,\ Aa) = \sum_{IBD=0,1,2} \Pr(Aa, Aa|IBD) * \Pr(IBD) = 4p^2q^2\pi_{0ij} + 4pq\phi_{ij}$$

and $\Pr(AA,\ aa) = 2p^2q^2\pi_{0ij}$. Therefore,

$$E(I_{Aa,Aa} - 2I_{AA,aa}) = \Pr(Aa, Aa) - 2\Pr(AA, aa) = 4\text{pq}\phi_{ij}$$

Supplementary Table S4.1: Probabilities for genotype pairs of bi-allelic variants given their IBD status.

	Pr(Genotype pairs|IBD)		
Genotype pairs	IBD=0	IBD=1	IBD=2
(aa,aa)	p_a^4	p_a^3	p_a^2
(aa,ab)	$2p_a^3p_b$	$p_a^2p_b$	0
(aa,bb)	$p_a^2p_b^2$	0	0
(ab,ab)	$4p_a^2p_b^2$	$p_a^2p_b + p_a p_b^2$	$2p_ap_b$

Suppose the genotype score, defined by the number of the reference allele for individuals i, is $X^{(i)}$.The absolute value of the genotype difference , $|X^{(i)} - X^{(j)}|$, takes three possible values: 2 with genotype pair (AA, aa), 1 with genotype pairs (AA, Aa) or (Aa, aa), and 0 otherwise. Thus we have

$$(X^{(i)} - X^{(j)})^2 = 4I_{AA,aa} + I_{Aa,aa} + I_{AA,Aa} = 4I_{AA,aa} + I_{Aa}{}^{(i)} + I_{Aa}{}^{(j)} - 2I_{Aa,Aa}$$

Under the assumption of HWE, we have $E\left(I_{Aa}^{(i)}\right) = \Pr(Aa) = 2pq$. It follows that

$$E(X^{(i)} - X^{(j)})^{\,2} = 4p(1-p)\left(1 - 2\phi_{ij}\right).$$

In the presence of population stratification, *P* may vary across individuals. To construct an estimator of kinship coefficient that is robust to population stratification, we assume P is a random variable representing AF of a randomly picked variant. P may vary across individuals but should follow the same distribution for a particular ancestry. Then the equation becomes

$$E\left(X^{(i)} - X^{(j)}\right)^2 = 4E\left(P(1-P)\right)\left(1 - 2\phi_{ij}\right).$$

Under the assumption of HWE, we have$\Pr(Aa|P) = 2P(1-P)$. Then

$$E\left(2P(1-P)\right) = E(\Pr(Aa|P)) = E\left(E(I_{Aa}|P)\right) = E(I_{Aa}).$$

For a pair of individuals, an empirical estimator for $E\left(2P(1-P)\right)$ is $(N_{Aa}^{(i)} + N_{Aa}^{(j)})/2M_{ij}$, where M_{ij} is the number of variants without missing value in both individuals.

The robust estimator for ϕ_{ij} is

$$\widehat{\phi_{ij}} = 0.5 - \frac{\sum_m (X_m^{(i)} - X_m^{(j)})^2}{2(N_{Aa}^{(i)} + N_{Aa}^{(j)})} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)} + N_{Aa}^{(j)}}$$

In the paper, a more robust estimator is proposed to deal with a situation when the violation of HWE of some variants results in excessive heterozygosity. Instead of using $(N_{Aa}^{(i)} + N_{Aa}^{(j)})/2M_{ij}$, they use $\min(N_{Aa}^{(i)}/M_{ij}, N_{Aa}^{(j)}/M_{ij})$ to estimate $E\left(2P(1-P)\right)$.

When the assumption of HWE is violated, the robust estimator for ϕ_{ij} is

$$\widehat{\phi_{ij}} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{2\min(N_{Aa}^{(i)}, N_{Aa}^{(j)})} + \frac{1}{2} - \frac{1}{4}\frac{N_{Aa}^{(i)} + N_{Aa}^{(j)}}{\min(N_{Aa}^{(i)}, N_{Aa}^{(j)})}$$

Once the kinship coefficient is calculated, relationship can be inferred based on criteria in Supplementary Table S4.2.

Supplementary Table S4.2: Relationship inference criteria for kinship coefficient.

Relationship	ϕ Inference criteria	π_0 Inference criteria
MZ twin	$>1/2^{3/2}$	<0.1
Parent-offspring	$(1/2^{5/2}, 1/2^{3/2})$	<0.1
Full-sib	$(1/2^{5/2}, 1/2^{3/2})$	(0.1,0.365)
Second degree relatives	$(1/2^{7/2}, 1/2^{5/2})$	(0.365,1-$1/2^{3/2}$)
Unrelated	$<1/2^{9/2}$	>1-$1/2^{5/2}$

Supplementary note 2: The influence of ignoring missing value information on the estimation of kinship coefficient.

To prevent study A from aggregating information of certain samples in study B, we proposed the strategy of permuting the sample order in each query (illustrated in Section 4.2.5). However, we have a concern that the missing information shared between studies may potentially reveal some identification information of samples of study B. To add another layer of protection on this protocol, study B should consider not sharing their missing information with study A. In other words, instead of calculating the kinship coefficient using $\widehat{\phi_{ij}} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)} + N_{Aa}^{(j)}}$, study A should approximate the kinship coefficient by

$$\widehat{\phi_{ij}} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)^*} + N_{Aa}^{(j)}},$$

where $N_{Aa}^{(i)^*}$ is calculated assuming study B has no missing value. Later, we call the kinship coefficient considering missing value in study B, the exact kinship coefficient; and the kinship coefficient ignoring missing value in study B, the approximate kinship coefficient.

Simulation was conducted on 1,000 selected TOPMed samples. These samples included 117 samples that consist of all the 108 pairs of duplicates, 400 samples having 1st degree relationship and 483 samples having 2nd degree relationship or being unrelated. The set of variants we used for relationship inference was exactly the same as the simulation in Section 4.3. The relationship was inferred by exact kinship coefficient as well as approximate kinship coefficient. We compared their relationship inference accuracy through true predicted value (TPV) and the false discovery rate (FDR). The TPV and FDR were calculated by comparing the results with the gold standard we described in Section 4.3.

Ignoring 0.2% missing value in real data

We did the simulation first based on the real TOPMed data. The missing rate across the variants per sample is about 0.2% for all the settings, 500, 1000, 5000 or 10000 variants.

The relationship inferred using the exact kinship coefficient is identical to using the approximate kinship coefficient under all the setting.

Ignoring 5% missing value in simulated data

In the first simulation, we show that not sharing missing value information of study B has no negative impact when the missing rate is low. Here we consider a situation where the missing rate is high. We simulated data by randomly masking the real data as missing with the probability of 0.05. Then relationship was inferred using the exact and approximate kinship coefficients.

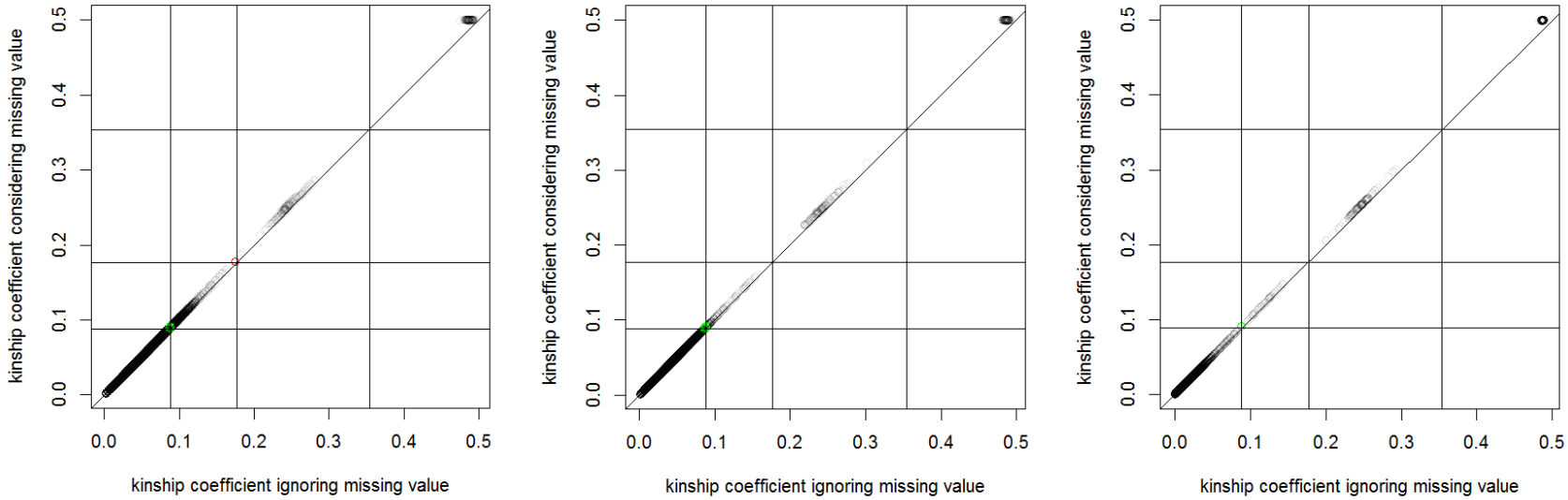
Results are shown in Supplementary Table S4.3. The results using approximate kinship coefficient are similar as those using exact kinship coefficient. With more than 5000 variants, the results are identical using these two kinship coefficients. As illustrated in Supplementary Figure S4.1, the approximate kinship coefficient underestimates the kinship coefficient because $N_{Aa}^{(i)^*}$ is larger than $N_{Aa}^{(i)}$. Relative pairs having kinship coefficient around the 0.354, 0.177 and 0.0884 cutoff are likely to be misclassified.

In conclusion, the impact of ignoring missing information of study B is negligible regardless of the missing rate. To infer relationship with our secure protocol, study B can keep its missing value information secret in order to protect its data from attacks.

Supplementary Table S4.3: Inference accuracy using exact kinship coefficient vs. using approximated kinship coefficient.

	Number of variants	TPV*	FDR**
Duplicates			
Exact kinship coefficient	500	100.00%	5.26%
	1000	100.00%	0.00%
	5000	100.00%	0.00%
	10000	100.00%	0.00%
Approx. kinship coefficient	500	100.00%	5.26%
	1000	100.00%	0.00%
	5000	100.00%	0.00%
	10000	100.00%	0.00%
1st degree relatives			
Exact kinship coefficient	500	85.59%	16.53%
	1000	86.36%	13.64%
	5000	84.55%	15.45%
	10000	85.05%	14.95%
Approx. kinship coefficient	500	85.59%	15.83%
	1000	86.36%	13.64%
	5000	84.55%	15.45%
	10000	85.05%	14.95%
2nd degree relatives			
Exact kinship coefficient	500	71.43%	97.90%
	1000	59.57%	82.61%
	5000	68.89%	38.00%
	10000	72.73%	25.58%
Approx. kinship coefficient	500	69.39%	97.57%
	1000	59.57%	77.95%
	5000	68.89%	36.73%
	10000	72.73%	25.58%

*TPV (true predictive value) = number of pairs inferred correctly/number of true pairs
**FDR(false discovery rate) = number of pairs inferred incorrectly/number of pairs inferred



Supplementary Figure S4.1: Kinship coefficients calculated by considering vs. ignoring missing information of study B. Grid corresponding to threshold of defining duplicates, 1st degree relatives and 2nd degree relatives (0.354, 0.177, 0.0884). Red dots are misclassified 1st degree relative pairs resulting from ignoring missing value. Green dots are misclassified 2nd degree relative pairs resulting from ignoring missing value.

Supplementary Table S4.4: Computational time of primary steps in the protocol under different parameter settings.

Number of variants	security	n	t	Encrypt/ entry	Evaluation overhead*	Evaluation/ pair	Decryption/ pair	Bandwidth/ pair
500	128	1024	10	1.26ms	16.08s	10.53ms	0.12ms	17kb
	128	1024	12	1.29ms	16.21s	7.21ms	0.14ms	17kb
	192	2048	10	2.45ms	30.73s	11.73ms	0.24ms	33kb
	192	2048	22	2.46ms	31.02s	13.61ms	0.24ms	33kb
1000	128	2048	12	2.51ms	63.57s	21.34ms	0.26ms	33kb
	128	2048	30	2.42ms	61.27s	21.05ms	0.25ms	33kb
	192	2048	12	2.34ms	59.50s	20.03ms	0.24ms	33kb
	192	2048	20	2.38ms	60.21s	21.01ms	0.21ms	33kb
5000	128	2048	14	2.55ms	320.78s	53.89ms	0.26ms	33kb
	128	2048	30	2.43ms	303.82s	59.06ms	0.23ms	33kb
	192	2048	14	2.15ms	269.02s	52.99ms	0.18ms	33kb
	192	2048	18	2.17ms	271.40s	55.32ms	0.19ms	33kb
10000	128	2048	16	2.52ms	633.49s	116.63ms	0.18ms	33kb
	128	2048	30	2.41ms	602.71s	103.00ms	0.18ms	33kb
	192	2048	16	2.26ms	565.98s	110.96ms	0.19ms	33kb
	192	2048	18	2.24ms	561.17s	108.45ms	0.20ms	33kb

* This step is to encrypt 25M zeros before the evaluation. Then 5M encrypted zeros is randomly selected to be added to the evaluation for function security. This step is only done one time by study B for all comparisons.

CHAPTER V

Summary and Future Work

5.1 Summary

Motivated by the genetic studies of electronic health records (EHR), this dissertation has focused on developing applicable methodologies to deal with the challenges of analyzing large-scale EHR data. Since EHR, originally, was not designed for scientific research, one critical drawback of using EHR for GWAS is the potential misclassification of phenotypes. Ignoring this misclassification will lead to biased association results and have a negative impact on the downstream analysis. Driven by this concern, we developed a method in Chapter II that can estimate the misclassification by examining external GWAS information. In addition, aggregating information between different EHR studies that have data sharing barriers requires privacy-preserving methods for relationship inference. This demand motivated the development of methods in Chapters III and IV. We proposed two secure protocols that can infer duplicates and genetic relatives between studies without sharing individual-level data.We believe these newly developed methods and protocols will facilitate analysis of genetic data along with EHR, and provide insight into future genetics research. However, each of the methods has both advantages and limitations, as will be described in this chapter. Moreover, they can be further improved to be more applicable to a broader range of problems. Therefore, we conclude this dissertation by pointing out potential directions for future research.

5.1.1 Modeling misclassification in phenotypes in EHR

One major challenge in EMR-based GWAS and PheWAS is the difficulty in accurately annotating disease phenotypes, which results from the low accuracy of billing codes as well as the difficulty of pooling billing codes to binary case/control phenotypes. Previous studies have demonstrated that ICD codes often have limited accuracy in predicting the true underlying disease status (Bazarian et al., 2006; Liao et al., 2010). In the analysis of the Michigan Genomic Initiative (MGI) data, we detected significant misclassification in age-related macular degeneration (AMD), psoriasis and type II diabetes (T2D) while no significant misclassification was detected in breast cancer. In Section 2.5, we noted that the misclassification can either occur along the translation from ICD codes to the dichotomized phenotypes or the assignment of ICD codes to patients. The possible errors in ICD codes introduced by O'Malley et al. (2005), such as the ambiguous description of ICD codes, the miscommunication between patients and clinicians, and the upcoding of ICD codes, all can explain the high misclassification we observe in MGI data. In addition, the observation by O'Malley et al., that a disease for which tests have high sensitivity will have higher diagnostic accuracy and smaller error in the ICD codes, explains the low misclassification of breast cancer in MGI. Our method provides researchers with guidance in tracing the origin of the errors and improving the case definition scheme.

Other than improving the phenotype construction, the estimation of our model, i.e. the misclassification rate, can be used to correct results in downstream analyses directly. First we can correct the effect size estimation with the misclassification rate using the formula derived by Neuhaus (1999) or Duffy's approach (2004), or using the iteratively reweighted least square algorithm (Magder and Hughes, 1997). In Chapter II, we corrected the effect size estimation for some variants associated with AMD, using Duffy's method, and saw the convergence of the

corrected results to the external GWAS. In the future, if applying the correction on the whole genome, we may recover the power of finding novel signals since the effect sizes are pulling against the null hypothesis. In addition, the misclassification rate can also be used to correct the receiver operating characteristic (ROC) curve analysis (Zawistowski et al.,2017).

5.1.2 Two relationship inference protocols

In Chapters III and IV, we proposed two protocols that can infer genetic relatives without compromising privacy. Although the methods were initially motivated by the demand of aggregating information between variant browsers, they can be applied to solve a broader range of problems when inferring relationships between studies is needed. Another promising direction is to use these methods for meta-analyses. Overlapping samples are found in meta-analysis when publicly available controls are shared among different studies (Young et al. 2007) or the same cohort contributes to different GWAS in a meta-analysis (Bonàs-Guarch et al. 2017). Overlapping samples can lead to inflated type I errors and false signals so that in case-control studies, the removal of related individuals is a standard quality control step (Voight and Pritchard, 2005). However, it can be challenging to remove overlapping samples in a meta-analysis, since individual-level data cannot be shared. Using our protocol, we will be able to identify the close relatives among different GWAS in a meta-analysis, then account for the overlaps using existing methods (Lin and Sullivan, 2009).

While the first protocol uses summary statistics to encrypt genotype data and uses a likelihood model to infer relationships, the second protocol uses homomorphic encryption to guarantee the security of the data and makes inference based on a robust method implemented in KING. They share some similarities in terms of utility. With 500 properly selected variants, both protocols have the ability to identify most of the 1st degree relatives with a low false discovery rate. With

5000 or more variants, 2nd degree relatives can be inferred potentially with low false discovery rate.

In addition to the similarities, both protocols have pros and cons compared with the other under different scenarios (Table 5.1). The first protocol is more computationally efficient. For the 10000 samples comparison, the first protocol takes less than 25 CPU hours while the best case for the second protocol is about 300 CPU hours, ignoring the communication overhead. The first protocol also scales well in bandwidth consumption. While the first protocol only requires MB-level of data communication, the second protocol requires hundreds of GB- or even TB- level of data communication overhead between studies.

However, the second protocol also has some obvious advantages. First of all, it can make robust inferences of relationships in a heterogeneous population. Especially when ancestry information is not available for selecting variants with consistent allele frequencies (AFs), the first protocol will have many false discoveries due to the biased AFs while the second protocol is robust to such bias. In addition, the second protocol encrypts genotypes with a higher level of security. Homomorphic encryption, as a well-established encryption scheme, has rigorous proof of the security level under different settings (Chase et al., 2017). The mapping between the encrypted value and the true value is totally random. Given the encrypted value, the corresponding true value can be any integers with equal probabilities. For the first protocol, on the other hand, the N-to-1 mapping between the encrypted value and the true value is fixed. Therefore given an encrypted value, the space to search for the true value is limited, which makes this protocol less secure than the second one.

In conclusion, the two methods are preferred under different scenarios. For inference in the homogenous population, one may consider using the method in Chapter III, which is more

efficient, while one may consider using the method in Chapter IV for the heterogeneous population and when higher level of security is required.

Table 5.1: Comparison between two relationship inference protocols proposed in Chapters III and IV.

	Chapter III	Chapter IV
Utility (Inference accuracy)*	European: TPV = 99.94%, FDR = 0.11%; Het. Pop.**: TPV = 100%, FDR = 0.15%	European: TPV = 99.93%, FDR = 0.07%; Het. Pop.: TPV = 100%, FDR = 0%
Security	**Summary statistics**: fixed N-to-1 mapping between the value before and after encryption	**HE**: random mapping between the value before and after encryption
Computation time***	25h	300h
Communication overhead***	6MB from A to B	324GB from A to B; 1621GB from B to A
Performance in heterogeneous population	Not robust, need to select variants with consistent AF across different populations	Robust

* Relationship inferred using 5000 variants
** Using selected 5000 variants based on AF
*** For 10000 vs. 10000 comparisons with 500 variants

5.2 Limitations and future work

One promising extension of the method in Chapter II is to deal with the misclassification between different cases and between cases and controls at the same time. For instance, it is very likely that T1D and T2D phenotypes are mixed in EHR because their ICD-9 code are quite similar (250.01, 250.03, 250.05, …, 250.93 for T1D and 250.00, 250.02, …, 250.92 for T2D) (Kho et al., 2011; Richesson et al., 2013). Other than considering additional information like

diagnostic lab tests to distinguish them, we can use a similar idea as what we proposed in Chapter II: examining genotype information of disease-associated variants. We can extend our model by incorporating another misclassification rate parameter that quantifies the misclassification between these two cases. Future development of methods dealing with misclassification between cases like T1D and T2D using genotype information will bring about more powerful and reliable genetic research using EHR data.

As discussed in Chapter II, our model is based on the assumption that the external GWAS is the gold-standard for the EHR GWAS. In other words, when the EHR and external GWAS have different underlying disease liability thresholds to dichotomize case/controls, those discrepancies due to the different threshold are also treated as a misclassification in our model. Taking psoriasis as an example, the GWAS of the purpose-built cohorts with dermatologist-diagnosed phenotypes usually only dichotomize the severe plaque type psoriasis as the case while the EHR GWAS treats patient with all kinds of psoriasis (Tsoi et al., 2017). Those differences are interpreted as misclassification in our model and cannot be separated from the true control-to-case misclassification. A method that can distinguish these two kinds of misclassification will be a promising future direction. It can benefit the interpretation of the results and help researchers to trace the source of misclassification.

Moreover, since the primary goal of Chapter II is to get accurate estimation of misclassification rate, we do not focus on the inference of the estimate. The mixed chi-square does not calibrate the likelihood ratio test statistic well under certain scenarios. A nonparametric bootstrap test has been shown in some research to have better performance than the asymptotic likelihood ratio test for testing parameters on the boundary (Cavaliere et al., 2017). In the future, we may consider

using this technique to control the inflated type I error and improve the inference of the misclassification rate.

For the two relationship inference protocols in Chapters III and IV, we did not evaluate their performance of identifying relatives of higher degrees, since our ultimate objective in this dissertation is to find the overlapping samples between studies. The methods, in the future, can be extended to infer any degree of relatives so that they can deal with a more general problem. Inspired by the method proposed by Epstein in 2000, we can build the likelihood-based protocol based on the hidden Markov model (HMM) considering the recombination probability between variants so that more precise relationships can be inferred. In addition, the single-nucleotide polymorphism microarray and the whole-genome sequencing enable more accurate detection of IBD segments and more precise resolution of IBD segment boundaries. Thus, multiple methods taking advantage of local IBD segment data have increased the range of detectable relationships up to 8^{th} degree (Huff et al., 2011; Li et al., 2014). By developing an encryption method for variants with higher density, we may achieve the goal of securely identifying higher degree relatives.

In order to infer higher degree relatives, more variants should be examined. The demand of incorporating more variants into the model generates another critical future direction of our protocols, improving computational efficacy. Sample sizes of recent research are exploding. For example, both the latest TOPMed and gnomAD freeze each have more than 100k samples and the UK Biobank, as described before, contains around 500k samples (Bycroft et al., 2018). As datasets continue to increase in size, we have to guarantee that our protocol is computationally feasible to deal with large-scale data. This can be overcome by taking advantage of the up-to-

date development of cryptography techniques, such as implementing more efficient homomorphic encryption techniques (Wang et al., 2018).

Last but not least, the protocol based on homomorphic encryption requires excessive communication bandwidth for large-scale studies. Both the degree of the polynomial and the coefficients are quite large and random for homomorphic encrypted ciphertext. As a consequence, an efficient way to store the ciphertext has been hard to develop. That is why there has not been much research in this area, even though this problem of homomorphic encryption is very common in real life settings. One potential solution is to hybrid homomorphic encryption with other encryption techniques that have much smaller ciphertext size, for example Advanced Encryption Standard (AES). The idea of the hybrid is that one study encrypts the data with AES rather than homomorphic encryption and then encrypts the AES key using homomorphic encryption, which means that calculation on the other side can be conducted by first decrypting the data from AES with encrypted AES key. Some studies have demonstrated that this strategy can solve certain kinds of practical problems in homomorphic encryption (Naehrig et al. 2011; Olumide et al. 2015; Alkady et al. 2018). We would like to address this important issue in the future by leveraging the knowledge accumulated through the development of this dissertation.

5.3 Closing remarks

Over the last decade, electronic health records have proved instrumental to unraveling the genetic complexities of disease risks. In particular, EHR-based GWAS and PheWAS are increasingly being used for locus replication as well as discovery of novel variants associated with complex diseases. The sequencing data being linked to EHR, in addition, expands the usage of EHR and helps with the functional interpretation of disease causing variants. One key advantage of using EHR is to boost study power by increasing the sample sizes of the association

study at low cost. However, challenges like misclassification in the phenotype as well as the concern of security when sharing data between studies need to be addressed.

Based upon this concept, by combining interdisciplinary knowledge of statistics, computer science, biology and medical science, my thesis first proposes a method to measure the misclassification in EHR-based GWAS and PheWAS analysis in order to reduce the bias in the search for disease-genetic associations and enhances the power of the analysis of the genomic basis of human disease. Then privacy preserving protocols are proposed to securely infer overlapping samples studies. It enables the accurate aggregation of information between studies, which not only can make the functional interpretation of putative disease-causing variants more precise but also can help avoiding spurious association results in meta-analysis. The continued expansion of GWAS, and its integration with the molecular functional interpretation, will be a critical asset for the study of gene coding and regulatory mechanisms and how they contribute to complex diseases. The subsequent downstream analysis that identifies biological pathways will provide information of suitable targets for drug development and repositioning of known therapeutics. Continuing steps toward filling the knowledge gap between genetics and disease will bring us closer to elucidating disease etiology and contribute to the development of preventive and improved treatment strategies.

BIBLIOGRAPHY

Alkady, Yasmin, Fifi Farouk, and Rawya Rizk. "Fully Homomorphic Encryption with AES in Cloud Computing Security." International Conference on Advanced Intelligent Systems and Informatics. Springer, Cham, 2018.

Ananthakrishnan, Ashwin N., et al. "Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach." Inflammatory bowel diseases 19.7 (2013): 1411.

Anderson, Amy D., and Bruce S. Weir. "A maximum-likelihood method for the estimation of pairwise relatedness in structured populations." Genetics 176.1 (2007): 421-440.

Ayday, Erman, Jean Louis Raisaro, and Jean-Pierre Hubaux. "Personal use of the genomic data: privacy vs. storage cost." Global Communications Conference (GLOBECOM), 2013 IEEE. IEEE, 2013.

Bajard, Jean-Claude, et al. "A full RNS variant of FV like somewhat homomorphic encryption schemes." International Conference on Selected Areas in Cryptography. Springer, Cham, 2016.

Bazarian, Jeffrey J., et al. "Accuracy of mild traumatic brain injury case ascertainment using ICD-9 codes." Academic emergency medicine 13.1 (2006): 31-38.

Bazerman, Max H., and William F. Samuelson. "I won the auction but don't want the prize." Journal of conflict resolution27.4 (1983): 618-634.

Blanton, Marina, et al. "Secure and efficient outsourcing of sequence comparisons." European Symposium on Research in Computer Security. Springer, Berlin, Heidelberg, 2012.

Boehnke, Michael, and Nancy J. Cox. "Accurate inference of relationships in sib-pair linkage studies." The American Journal of Human Genetics 61.2 (1997): 423-429.

Bonàs-Guarch, Sílvia, et al. "A comprehensive reanalysis of publicly available GWAS datasets reveals an X chromosome rare regulatory variant associated with high risk for type 2 diabetes." bioRxiv (2017): 112219.

Bringer, Julien, Hervé Chabanne, and Alain Patey. "Shade: Secure hamming distance computation from oblivious transfer." International Conference on Financial Cryptography and Data Security. Springer, Berlin, Heidelberg, 2013.

Broman, Karl W., and James L. Weber. "Estimation of pairwise relationships in the presence of genotyping errors." American journal of human genetics 63.5 (1998): 1563.

Bush, William S., Matthew T. Oetjens, and Dana C. Crawford. "Unravelling the human genome-phenome relationship using phenome-wide association studies." Nature Reviews Genetics 17.3 (2016): 129-145.

Bycroft, Clare, et al. "The UK Biobank resource with deep phenotyping and genomic data." Nature 562.7726 (2018): 203.

Carroll, Robert J., et al. "Portability of an algorithm to identify rheumatoid arthritis in electronic health records." Journal of the American Medical Informatics Association 19.e1 (2012): e162-e169.

Carroll, Robert J., Lisa Bastarache, and Joshua C. Denny. "R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment." Bioinformatics30.16 (2014): 2375-2376.

Casella, George; Berger, Roger L. (2001). Statistical Inference (Second ed.). ISBN 0-534-24312-6.

Cavaliere, Giuseppe, Heino Bohn Nielsen, and Anders Rahbek. "On the Consistency of Bootstrap Testing for a Parameter on the Boundary of the Parameter Space." Journal of Time Series Analysis 38.4 (2017): 513-534.

Cavalli-Sforza, L. Luca. "The human genome diversity project: past, present and future." Nature Reviews Genetics 6.4 (2005): 333.

Chase, Melissa, et al. "Security of homomorphic encryption." HomomorphicEncryption. org, Redmond WA, Tech. Rep(2017).

Chen, Hao, et al. "Simple encrypted arithmetic library v2. 3.0." Microsoft Research TechReport (2017).

Cheon, Jung Hee, Miran Kim, and Kristin Lauter. "Homomorphic computation of edit distance." International Conference on Financial Cryptography and Data Security. Springer, Berlin, Heidelberg, 2015.

Consortium, Diabetes SAT2D, et al. "Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility." Nature genetics46.3 (2014): 234.

Copeland, Karen T., et al. "Bias due to misclassification in the estimation of relative risk." American journal of epidemiology 105.5 (1977): 488-495.

De Cristofaro, Emiliano, and Gene Tsudik. "Experimenting with fast private set intersection." International Conference on Trust and Trustworthy Computing. Springer, Berlin, Heidelberg, 2012.

Denny, Joshua C., et al. "PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations." Bioinformatics26.9 (2010): 1205-1210.

Denny, Joshua C., et al. "Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data." Nature biotechnology 31.12 (2013): 1102-1111.

Duffy, S. W., et al. "A simple model for potential use with a misclassified binary outcome in epidemiology." Journal of Epidemiology & Community Health 58.8 (2004): 712-717.

Epstein, Michael P., William L. Duren, and Michael Boehnke. "Improved inference of relationship for pairs of individuals." The American Journal of Human Genetics 67.5 (2000): 1219-1231.

Fan, Junfeng, and Frederik Vercauteren. "Somewhat Practical Fully Homomorphic Encryption." IACR Cryptology ePrint Archive 2012 (2012): 144.

Fletcher, Olivia, et al. "Novel breast cancer susceptibility locus at 9q31. 2: results of a genome-wide association study." Journal of the National Cancer Institute 103.5 (2011): 425-435.

Fritsche, Lars G., et al. "A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants." Nature genetics 48.2 (2016): 134.

Gaulton, Kyle J., et al. "Genetic fine-mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci." Nature genetics 47.12 (2015): 1415.

Gentry, Craig. A fully homomorphic encryption scheme. Stanford University, 2009.

Gordon, Derek, et al. "Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double-sampling." Statistical applications in genetics and molecular biology 3.1 (2004): 1-32.

Haplotype Reference Consortium. "A reference panel of 64,976 haplotypes for genotype imputation." Nature genetics 48.10 (2016): 1279-1283.

He, Dan, et al. "Identifying genetic relatives without compromising privacy." Genome research 24.4 (2014): 664-672.

Holm, Sture. "A simple sequentially rejective multiple test procedure." Scandinavian journal of statistics (1979): 65-70.

Homer, Nils, et al. "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays." PLoS genetics 4.8 (2008): e1000167.

Hong, Chuan, Katherine P. Liao, and Tianxi Cai. "Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping." Biometrics 75.1 (2019): 78-89.

Hormozdiari, Farhad, et al. "Privacy preserving protocol for detecting genetic relatives using rare variants." Bioinformatics30.12 (2014): i204-i211.

Huff, Chad D., et al. "Maximum-likelihood estimation of recent shared ancestry (ERSA)." Genome research 21.5 (2011): 768-774.

Illumina (2017). Infinium® CoreExome-24 v1.2 BeadChip, San Diego, CA.

Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." Nature reviews. Genetics 13.6 (2012): 395.

Kantarcioglu, Murat, et al. "A cryptographic approach to securely share and query genomic sequences." IEEE Transactions on information technology in biomedicine 12.5 (2008): 606-617.

Karczewski, Konrad J., et al. "Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes." BioRxiv(2019): 531210.

Kho, Abel N., et al. "Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study." Journal of the American Medical Informatics Association 19.2 (2011): 212-218.

Kim, Miran, and Kristin Lauter. "Private genome analysis through homomorphic encryption." BMC medical informatics and decision making. Vol. 15. No. 5. BioMed Central, 2015.

Landrum, Melissa J., et al. "ClinVar: improving access to variant interpretations and supporting evidence." Nucleic acids research 46.D1 (2017): D1062-D1067.

Lange K, Optimization, 2004, Springer.

Lauter, Kristin, Adriana López-Alt, and Michael Naehrig. "Private computation on encrypted genomic data." International Conference on Cryptology and Information Security in Latin America. Springer, Cham, 2014.

Lek, Monkol, et al. "Analysis of protein-coding genetic variation in 60,706 humans." Nature 536.7616 (2016): 285.

Li, Hong, et al. "Relationship estimation from whole-genome sequence data." PLoS genetics 10.1 (2014): e1004144.

Liao, Katherine P., et al. "Electronic medical records for discovery research in rheumatoid arthritis." Arthritis care & research 62.8 (2010): 1120-1127.

Lin, Dan-Yu, and Patrick F. Sullivan. "Meta-analysis of genome-wide association studies with overlapping subjects." The American Journal of Human Genetics 85.6 (2009): 862-872.

Lohmueller, Kirk E., et al. "Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease." Nature genetics 33.2 (2003): 177.

Lynch, Michael. "Estimation of relatedness by DNA fingerprinting." Molecular biology and evolution 5.5 (1988): 584-599.

Lynch, Michael, and Kermit Ritland. "Estimation of pairwise relatedness with molecular markers." Genetics 152.4 (1999): 1753-1766.

Madsen K, Nielsen HB, Tingleff O, Optimization With Constraints, 2004, IMM, Technical University of Denmark.

Magder, Laurence S., and James P. Hughes. "Logistic regression when the outcome is measured with uncertainty." American Journal of Epidemiology146.2 (1997): 195-203.

Mailman, Matthew D., et al. "The NCBI dbGaP database of genotypes and phenotypes." Nature genetics 39.10 (2007): 1181.

Manichaikul, Ani, et al. "Robust relationship inference in genome-wide association studies." Bioinformatics 26.22 (2010): 2867-2873.

Martin R. Albrecht, Rachel Player, and Sam Scott. On the concrete hardness of learning with errors. J. Mathematical Cryptology, 9(3):169–203, 2015.

McCarty, Catherine A., et al. "The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies." BMC medical genomics 4.1 (2011): 13.

McDavid, Andrew, et al. "Enhancing the power of genetic association studies through the use of silver standard cases derived from electronic medical records." PloS one 8.6 (2013): e63481.

Michailidou, Kyriaki, et al. "Large-scale genotyping identifies 41 new loci associated with breast cancer risk." Nature genetics 45.4 (2013): 353.

Michailidou, Kyriaki, et al. "Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer." Nature genetics 47.4 (2015): 373.

Michailidou, Kyriaki, et al. "Association analysis identifies 65 new breast cancer risk loci." Nature 551.7678 (2017): 92.

Michigan Genome Initiative https://precisionhealth.umich.edu/michigangenomics/

Milligan, Brook G. "Maximum-likelihood estimation of relatedness." Genetics 163.3 (2003): 1153-1167.

Morris, Andrew P., et al. "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes." Nature genetics 44.9 (2012): 981.

Naehrig, Michael, Kristin Lauter, and Vinod Vaikuntanathan. "Can homomorphic encryption be practical?." Proceedings of the 3rd ACM workshop on Cloud computing security workshop. ACM, 2011.

Neuhaus, John M. "Bias and efficiency loss due to misclassified responses in binary regression." Biometrika 86.4 (1999): 843-855.

Nielsen, Jonas B., et al. "Biobank-driven genomic discovery yields new insight into atrial fibrillation biology." Nature genetics 50.9 (2018): 1234.

O'malley, Kimberly J., et al. "Measuring diagnoses: ICD code accuracy." Health services research 40.5p2 (2005): 1620-1639.

Olumide, Atewologun, et al. "A hybrid encryption model for secure cloud computing." 2015 13th International Conference on ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015). IEEE, 2015.

PheWeb, http://pheweb.sph.umich.edu/

Queller, David C., and Keith F. Goodnight. "Estimating relatedness using genetic markers." Evolution 43.2 (1989): 258-275.

Richesson, Rachel L., et al. "A comparison of phenotype definitions for diabetes mellitus." Journal of the American Medical Informatics Association20.e2 (2013): e319-e326.

Risch, Neil. "Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs." American journal of human genetics 46.2 (1990): 242.

Ritchie, M.D. et al. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. Circulation 127, 1377–1385 (2013).

Ritland, Kermit. "Estimators for pairwise relatedness and individual inbreeding coefficients." Genetics Research 67.2 (1996): 175-185.

Saxena, Richa, et al. "Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci." The American Journal of Human Genetics 90.3 (2012): 410-425.

Scott, Robert A., et al. "An expanded genome-wide association study of type 2 diabetes in Europeans." Diabetes66.11 (2017): 2888-2902.

Self, Steven G., and Kung-Yee Liang. "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions." Journal of the American Statistical Association 82.398 (1987): 605-610.

Sherry, Stephen T., et al. "dbSNP: the NCBI database of genetic variation." Nucleic acids research 29.1 (2001): 308-311.

Shimizu, Kana, Koji Nuida, and Gunnar Rätsch. "Efficient privacy-preserving string search and an application in genomics." Bioinformatics 32.11 (2016): 1652-1661.

Sinnott, Jennifer A., et al. "Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records." Human genetics 133.11 (2014): 1369-1382.

Strange, Amy, et al. "A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1." Nature genetics 42.11 (2010): 985.

Stuart, Philip E., et al. "Genome-wide association analysis identifies three psoriasis susceptibility loci." Nature genetics42.11 (2010): 1000.

Sudlow, Cathie, et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." PLoS medicine 12.3 (2015): e1001779.

Taliun, Daniel, et al. "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program." BioRxiv (2019): 563866.

Thomas, Stuart C. "A simplified estimator of two and four gene relationship coefficients." Molecular ecology resources 10.6 (2010): 986-994.

Thompson, E. A. "The estimation of pairwise relationships." Annals of human genetics 39.2 (1975): 173-188.

Tsoi, Lam C., et al. "Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity." Nature genetics 44.12 (2012): 1341-1348.

Tsoi, Lam C., et al. "Enhanced meta-analysis and replication studies identify five new psoriasis susceptibility loci." Nature communications 6 (2015): 7001.

Tsoi, Lam C., et al. "Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants." Nature communications 8 (2017): 15382.

Turnbull, Clare, et al. "Genome-wide association study identifies five new breast cancer susceptibility loci." Nature genetics 42.6 (2010): 504.

Ugwuoke, Chibuike, Zekeriya Erkin, and Reginald L. Lagendijk. "Privacy-safe linkage analysis with homomorphic encryption." 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017.

Uhlerop, Caroline, Aleksandra Slavković, and Stephen E. Fienberg. "Privacy-preserving data sharing for genome-wide association studies." The Journal of privacy and confidentiality5.1 (2013): 137.

UK Biobank. (2018). Regeneron announces major collaboration to exome sequence UK Biobank genetic data more quickly. Retrieved from http://www.ukbiobank.ac.uk/2018/01/regeneron-announces-major-collaboration-to-exome-sequence-uk-biobank-genetic-data-more-quickly/; date last accessed March 14, 2018.

Voight, Benjamin F., and Jonathan K. Pritchard. "Confounding from cryptic relatedness in case-control association studies." PLoS genetics 1.3 (2005): e32.

Voight, Benjamin F., et al. "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis." Nature genetics 42.7 (2010): 579.

Wang, Jinliang. "An estimator for pairwise relatedness using molecular markers." Genetics 160.3 (2002): 1203-1215.

Wang, J. I. N. L. I. A. N. G. "Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods." Journal of Evolutionary Biology 27.3 (2014): 518-530.

Wang, Xiao Shaun, et al. "Efficient genome-wide, privacy-preserving similar patient query based on private edit distance." Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, 2015.

Wang, Chaolong, et al. "Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation." The American Journal of Human Genetics 96.6 (2015): 926-937.

Wang, J. "Estimating pairwise relatedness in a small sample of individuals." Heredity 119.5 (2017): 302.

Wang, Xun, Tao Luo, and Jianfeng Li. "A More Efficient Fully Homomorphic Encryption Scheme Based on GSW and DM Schemes." Security and Communication Networks 2018 (2018).

Weissbrod, Omer, et al. "Accurate liability estimation improves power in ascertained case-control studies." Nature methods 12.4 (2015): 332-334.

Wilke RA, Berg RL, Vidaillet HJ, Caldwell MD, Burmester JK, Hillman MA. Impact of age, CYP2C9 genotype and concomitant medication on the rate of rise for prothrombin time during the first 30 days of warfarin therapy. Clin Med Res. 2005;3:207–213.

Wolford, Brooke N., Cristen J. Willer, and Ida Surakka. "Electronic health records: the next wave of complex disease genetics." Human molecular genetics 27.R1 (2018): R14-R21.

Yin, Xianyong, et al. "Genome-wide meta-analysis identifies multiple novel associations and ethnic heterogeneity of psoriasis susceptibility." Nature communications 6 (2015): 6916.

Young, A. H., et al. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature (2007).

Yu, Fei, et al. "Scalable privacy-preserving data sharing methodology for genome-wide association studies." Journal of biomedical informatics 50 (2014): 133-141.

Zawistowski, Matthew, et al. "Corrected ROC analysis for misclassified binary outcomes." Statistics in medicine 36.13 (2017): 2148-2160.

Zeggini, Eleftheria, et al. "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." Nature genetics 40.5 (2008): 638.

Zhou, Wei, et al. "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies." Nature genetics 50.9 (2018): 1335.