2023

Statistical methods for gene selection and genetic association studies

Xuewei Cao
Michigan Technological University, xueweic@mtu.edu

STATISTICAL METHODS FOR GENE SELECTION AND GENETIC ASSOCIATION STUDIES

By

Xuewei Cao

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Statistics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2023

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Statistics.

Department of Mathematical Sciences

Dissertation Advisor:	*Dr. Qiuying Sha*
Committee Member:	*Dr. Shuanglin Zhang*
Committee Member:	*Dr. Kui Zhang*
Committee Member:	*Dr. Weihua Zhou*
Department Chair:	*Dr. Jiguang Sun*

Table of Contents

Author Contribution Statement

This dissertation is submitted for the degree of Doctor of Philosophy at Michigan Technological University. The research represented in this dissertation was conducted under the supervision of professor Qiuying Sha in the Department of Mathematical Sciences, Michigan Technological University, between August 2018 and April 2023. This dissertation contains published, completed papers, some important preparations, and achievements for future publications completed by the author. This work is to the best of my knowledge original, except where references are made to previous work.

The first chapter, *A novel method for multiple phenotype association studies based on genotype and phenotype network*, was submitted to Genome Research in February 2023. The overall study was designed by Dr. Qiuying Sha and Dr. Shuanglin Zhang. Xuewei Cao performed statistical analyses, interpreted the results through data curation and visualization, built the public software, wrote the original manuscript under the supervision of Dr. Qiuying Sha and Dr. Shuanglin Zhang.

The second chapter, *Constructing genotype and phenotype network helps reveal disease heritability and phenome-wide association studies*, is in preparation for future publication. The overall study was designed by Xuewei Cao under the supervision of Dr. Qiuying Sha and Dr. Shuanglin Zhang. Xuewei Cao and Lirong Zhu performed network constructions, statistical analyses, real date applications in collaboration with Dr. Xiaoyu Liang from Department of Epidemiology and Biostatistics at Michigan State University.

The third chapter, *Gene-based association tests using GWAS summary statistics and incorporating eQTL*, was published in Scientific Reports in 2022. Xuewei Cao led this work and performed the statistical analyses, under the supervision of Dr. Qiuying Sha and Dr. Shuanglin Zhang, in collaboration with Dr. Xuewia Wang from Department of Biostatistics at Florida International University.

The fourth chapter, *TGPred: Efficient methods for predicting target genes of a transcription factor by integrating statistics, machine learning, and optimization*, was submitted to NAR Genomics and Bioinformatics in November 2022, which is in revision. Xuewei Cao and Ling Zhang led this work, performed the statistical analyses, wrote the original manuscript, and built the public software in R and Python under the supervision of Dr. Qiuying Sha and Dr. Hairong Wei. This work was also collaborated with Dr. Kui Zhang from Department of Mathematical Sciences at Michigan Technological University, Dr. Sanzhen Liu, Dr. Mingxia Zhao, and Dr. Cheng He from Department of Plant Pathology at Kansas State University. All collaborators provided advice during the research progress, checked the technical works, and edited the manuscript.

The fifth chapter, *Gene selection by incorporating genetic networks into case-control association studies*, was published in European Journal of Human Genetics in 2022. Xuewei Cao led this work and performed the statistical analyses, under the supervision of Dr. Qiuying Sha and Dr. Shuanglin Zhang, in collaboration with Dr. Xiaoyu Liang from Department of Epidemiology and Biostatistics at Michigan State University.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Qiuying Sha, for all of her valuable guidance, continuous support, and encouragement throughout the research process over the last five years, my entire Ph.D. study. Without her guidance and support, I would not have been able to achieve the level of understanding and skill that I have today. Her patience, motivation, enthusiasm, and immense insights have been invaluable in shaping this dissertation. Dr. Sha serves not only as an academic advisor but also as a mentor to my personality. I cannot ask for a better advisor.

Besides my advisor, I would also like to thank my committee members, Professor Shuanglin Zhang, Professor Kui Zhang, and Professor Weihua Zhou, for their insightful comments and for dedicating their time to review and provide input on my dissertation. In particular, Professor Shuanglin Zhang's expertise, mentorship, and willingness to answer my questions have been instrumental in the development of this dissertation.

Special thanks to my friends and collaborators who provided me with great help and support. I would not be able to name all of them, but I must thank Dr. Yanfang Liu, Lirong Zhu, and Meida Wang for their support and friendly to my life at MTU. After five years at Michigan Technological University, I have been deeply loving this place because of all the lovely people.

Thanks all the funding resources for the supports, which allows me to be able to concentrate all of my time and efforts on my research projects, including Portage Health Foundation (PHF) Graduate Assistantship, Health Research Institute Graduate Fellowship, Doctoral Finishing Fellowship of Michigan Technological University, and the NIH R15 and PHF endowed professor grants from my advisor.

Last but not least, I am extremely grateful to my family, especially my parents, Yongsheng Cao and Zhihua Liu, for their love, caring, and sacrifices in educating and preparing me for my future, so that I can reach this stage.

Once again, I am deeply grateful for the support and guidance of everyone who has contributed to this dissertation.

Abstract

This dissertation includes five Chapters. A brief description of each chapter is organized as follows.

In Chapter One, we propose a signed bipartite genotype and phenotype network (GPN) by linking phenotypes and genotypes based on the statistical associations. It provides a new insight to investigate the genetic architecture among multiple correlated phenotypes and explore where phenotypes might be related at a higher level of cellular and organismal organization. We show that multiple phenotypes association studies by considering the proposed network are improved by incorporating the genetic information into the phenotype clustering.

In Chapter Two, we first illustrate the proposed GPN to GWAS summary statistics. Then, we assess contributions to constructing a well-defined GPN with a clear representation of genetic associations by comparing the network properties with a random network, including connectivity, centrality, and community structure. The network topology annotations based on the sparse representations of GPN can be used to understand the disease heritability for the highly correlated phenotypes. In applications of phenome-wide association studies, the proposed GPN can identify more significant pairs of genetic variant and phenotype categories.

In Chapter Three, a powerful and computationally efficient gene-based association test is proposed, aggregating information from different gene-based association tests and also incorporating expression quantitative trait locus information. We show that the proposed method controls the type I error rates very well and has higher power in the simulation studies and can identify more significant genes in the real data analyses.

In Chapter Four, we develop six statistical selection methods based on the penalized regression for inferring target genes of a transcription factor (TF). In this study, the proposed selection methods combine statistics, machine learning , and convex optimization approach, which have great efficacy in identifying the true target genes. The methods will fill the gap of lacking the appropriate methods for predicting target genes of a TF, and are instrumental for validating experimental results yielding from ChIP-seq and DAP-seq, and conversely, selection and annotation of TFs based on their target genes.

In Chapter Five, we propose a gene selection approach by capturing gene-level signals in network-based regression into case-control association studies with DNA sequence data or DNA methylation data, inspired by the popular gene-based association tests using a weighted combination of genetic variants to capture the combined effect of individual genetic variants within a gene. We show that the proposed gene selection approach have higher true positive rates than using traditional dimension reduction techniques in the simulation studies and select potentially rheumatoid arthritis related genes that are missed by existing methods.

1 Chapter 1

A novel method for multiple phenotype association studies based on genotype and phenotype network

Abstract

Joint analysis of multiple correlated phenotypes for genome-wide association studies (GWAS) can identify and interpret pleiotropic loci which are essential to understand pleiotropy in diseases and complex traits. Meanwhile, constructing a network based on associations between phenotypes and genotypes provides a new insight to analyze multiple phenotypes, which can explore whether phenotypes and genotypes might be related to each other at a higher level of cellular and organismal organization. In this paper, we first develop a bipartite signed network by linking phenotypes and genotypes into a Genotype and Phenotype Network (GPN). The GPN can be constructed by a mixture of quantitative and qualitative phenotypes and is applicable to binary phenotypes with extremely unbalanced case-control ratios in large-scale biobank datasets. We then apply a powerful community detection method to partition phenotypes into disjoint network modules based on GPN. Finally, we jointly test the association between multiple phenotypes in a network module and a single nucleotide polymorphism (SNP). Simulations and analyses of 72 complex traits in the UK Biobank show that multiple phenotype association tests based on network modules detected by GPN are much more powerful than those without considering network modules. The newly proposed GPN provides a new insight to investigate the genetic architecture among different types of phenotypes. Multiple phenotypes association studies based on GPN are improved by incorporating the genetic information into the phenotype clustering. Notably, it might broaden the understanding of genetic architecture that exists between diagnoses, genes, and pleiotropy.

Keywords: multiple phenotype association studies; genotype and phenotype network; community detection

1.1 Introduction

Genome-wide association studies (GWAS) have successfully identified thousands of single nucleotide polymorphisms (SNPs) genetically associated with a wide range of complex human diseases and traits[1,2]. Over the past decade, more than 10,000 associations between SNPs and diseases/traits have been discovered[3]. Although GWAS have emerged as a common and powerful tool to detect the complexity of the genotype-phenotype associations, a common limitation of GWAS is that they focus on only a single phenotype at a time[4-7]. Joint analysis of multiple correlated phenotypes for GWAS may provide more power to identify and interpret pleiotropic loci, which are essential to understand pleiotropy in diseases and complex traits[4,8,9]. In brief, biological pleiotropy refers to a SNP or gene that has a direct biological influence on more than one phenotypic trait[10]. Biological

pleiotropy can offer significant insights in understanding the complex genotype-phenotype relationships[2]. Therefore, multiple phenotypes are usually collected in many GWAS cohorts and jointly analyzing multiple phenotypes may increase statistical power to discover the cross-phenotype associations and pleiotropy[10-13].

Many statistical methods have been developed to jointly test the association between a SNP and multiple correlated phenotypes[14]. The most widely used methods for multiple phenotype association studies can be roughly classified into three categories: 1) statistical tests based on combining either the univariate test statistics or p-values, such as O'Brien's method[15], adaptive Fisher's combination (AFC)[16], aSPU[17], and others[18]; 2) multivariate analyses based on regression methods, such as multivariate analysis of variance (MANOVA)[19], reverse regression methods (MultiPhen)[20], linear mixed effect models (LMM)[21], and generalized estimating equations (GEE)[22]; and 3) dimension reduction methods, such as clustering linear combination (CLC)[12], canonical correlation analysis (CCA)[23], and principal components analysis (PCA)[24,25]. However, most phenotypes are influenced by many SNPs that act in concert to alter cellular function[26], the above mentioned methods are only based on phenotypic correlation without considering the genetic correlation among phenotypes. Therefore, these methods may loss statistical power to detect the true pleiotropic effects comparing the methods based on genetic architecture among complex diseases. To address this issue, numerous types of algorithms to investigate the genetic correlation among complex traits and diseases have been developed[27-29]. Many of these algorithms are often in conjunction with linkage disequilibrium (LD) information by using GWAS summary association data[28]. For example, cross-trait LD score regression has been developed to estimate genetic and phenotypic correlation that requires only GWAS summary statistics and is not biased by overlapping samples[27].

In 2007, a conceptually different approach based on the human disease network had been developed, exploring whether human complex traits and the corresponding genotypes might be related to each other at a higher level of cellular and organismal organization[30]. Network analyses provide an integrative approach to characterize complex genomic associations[31]. Therefore, constructing a network based on the associations between phenotypes and genotypes provides a new insight to simultaneously analyze multiple phenotypes and SNPs. Notably, it might broaden the understanding of genetic architecture that exists between diagnoses, genes, and pleiotropy[8]. Modules detected from human disease networks are useful in providing insights pertaining to biological functionality[32]. Therefore, community detection methods play a key role in understanding the global and local structures of disease interaction, in shedding light on association connections that may not be easily visible in the network topology[33]. Many community detection methods have been applied from social networks to human disease networks, such as Louvain's method[8] with modularity as a measure and core module identification to identify small and structurally well-defined communities[32]. However, most community detection methods have been developed for unsigned networks[34-40].

To date, many biobanks, such as the UK Biobank[41], aggregate data across tens of thousands of phenotypes and provide a great opportunity to construct the human disease network and perform joint analyses of multiple correlated phenotypes. The electronic

health record (EHR)-driven genomic research (EDGR) workflow is the most popular way to analyze multiple diagnosis codes in Biobank data, at its core, which is the use of EHR data for genomic research in the investigation of population-wide genomic characterization[42]. In most EHR systems, the whole phenome can be divided into numerous phenotypic categories according to the first few characters of the International Classification of Disease (ICD) billing codes[43]. However, the ICD-based categories are based on the underlying cause of death rather than on the shared genetic architecture among all complex diseases and traits. Meanwhile, the phenotypes in large biobanks usually have extremely unbalanced case-control ratios. Therefore, linking phenotypes, especially EHR-derived phenotypes, with genotypes in a network is also very important to examine the genetic architecture of complex diseases and traits.

1.2 Material and Methods

1.2.1 Overview of Methods

In this paper, we develop a bipartite signed network by linking phenotypes and genotypes into a Genotype and Phenotype Network (GPN; **Figure 1.1a**). The GPN can be constructed by a mixture of quantitative and qualitative phenotypes and is applicable to phenotypes with extremely unbalanced case-control ratios for large-scale biobank datasets since the saddlepoint approximation[44] is used to test the association between genotype and phenotype with extremely unbalanced case-control ratio. After projecting genotypes into phenotypes, the genetic correlation of phenotypes can be calculated based on the shared associations among all genotypes (**Figure 1.1b**). We then apply a powerful community detection method to partition phenotypes into disjoint network modules using the hierarchical clustering method and the number of modules is determined by perturbation (**Figure 1.1c**)[45]. The phenotypes in each network module share the same genetic information. After partitioning phenotypes into disjoint network modules, a statistical method for multiple phenotype association studies can be applied to test the association between phenotypes in each module and a SNP, then a Bonferroni correction can be used to test if all phenotypes are associated with a SNP (**Figure 1.1d**). To jointly analyze the association between multiple phenotypes in each module with a SNP, we use six multiple phenotype association tests, including ceCLC[46], CLC[12], HCLC[47], MultiPhen[20], O'Brien[15], and Omnibus[12]. The advantage of the association test based on network modules detected by GPN is that phenotypes in a network module are highly correlated based on the genetic architecture, therefore, the association test is more powerful to identify pleiotropic SNPs. After we obtain the GWAS signals from the previous steps, post-GWAS analyses can be applied to understand the high level of biological mechanism, such as pathway/tissue enrichment analysis and colocalization of GWAS signals and eQTL analysis in the specific disease-associated tissue (**Figure 1.1e-g**). The construction of GPN, community detection method, and six multiple phenotype association tests with and without considering the network modules detected by GPN have been implemented in R, which is an open-source software and publicly available on GitHub: https://github.com/xueweic/GPN.

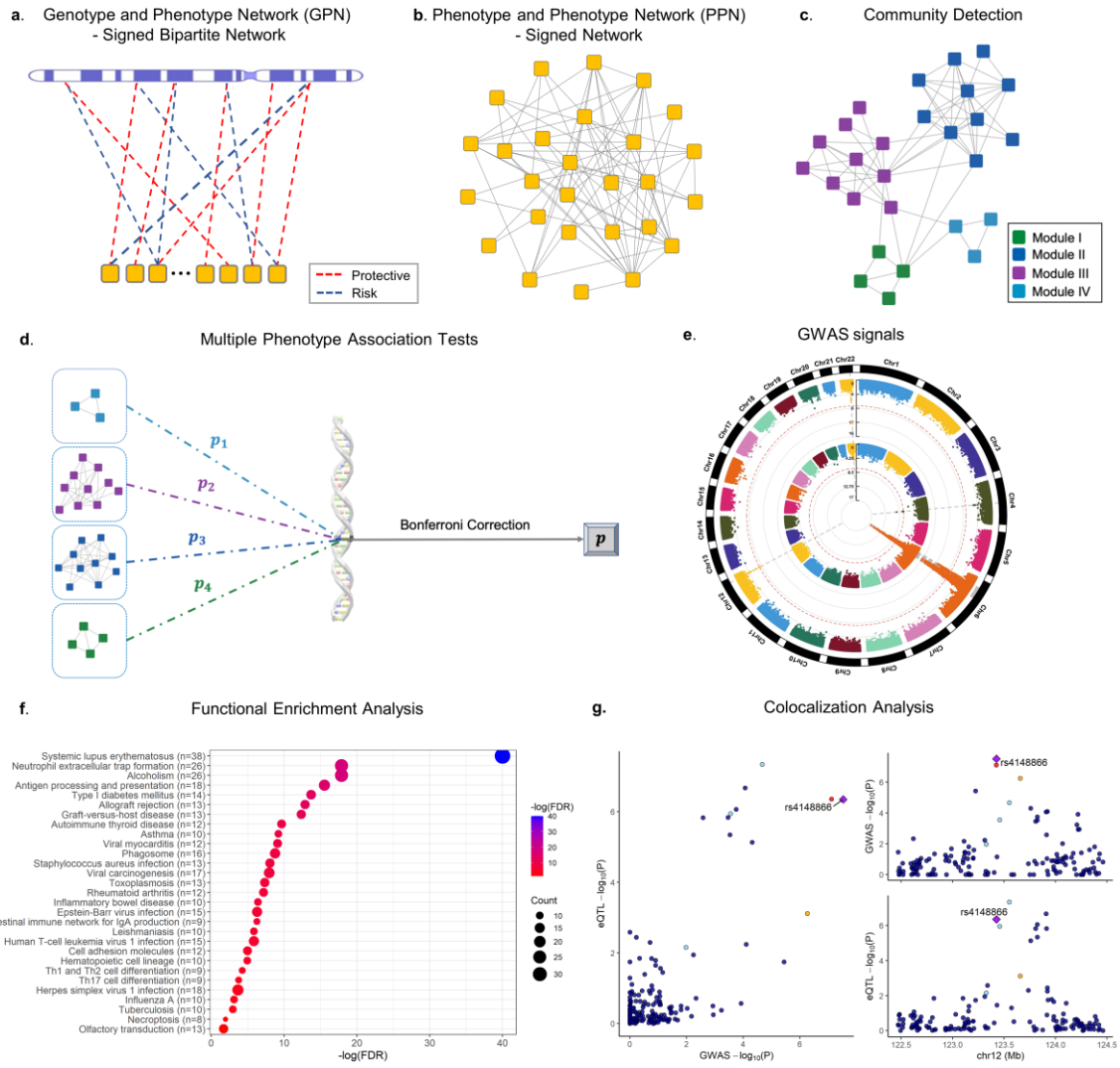


Figure 1.1. Overview of the method. **a**. Construction of a signed bipartite network, GPN. Each phenotype (yellow square) and each SNP form a directed edge which represents the strength of the association, where the red dashed line indicates that the minor allele of the SNP is a protective allele to the phenotype, and the blue dashed line indicates that the minor allele of the SNP is a risk allele to the phenotype. **b**. Construction of a signed network, PPN, which is the one-mode projection of GPN on phenotypes. **c**. The powerful community detection method is used to partition phenotypes into disjoint network modules with different colors. **d**. Multiple phenotype association tests are applied to test the association between phenotypes in each of the network modules and a SNP, then the Bonferroni correction is used to obtain the overall p-value. **e**. GWAS signals are identified by a multiple phenotype association test with or without considering network modules. **f**. Functional enrichment analysis based on the detected GWAS signals and the publicly available functional database. **g**. Colocalization of GWAS signals and eQTL analysis.

1.2.2 Construction of the Genotype and Phenotype Network

Consider a sample with n unrelated individuals, indexed by $i=1,\cdots,n$. Suppose each individual has a total of K phenotypes and M SNPs. Let $\mathbf{Y}=(y_{ik})$ be an $n\times K$ matrix of K phenotypes, where y_{ik} denotes the phenotype value of the i^{th} individual for the k^{th} phenotype. The phenotypes can be both quantitative and qualitative, especially for phenotypes with extremely unbalanced case-control ratios. Let $\mathbf{G}=(g_{im})$ be an $n\times M$ matrix of genotypes, where g_{im} represents the genotypic score of the i^{th} individual at the m^{th} SNP which is the number of minor alleles that the i^{th} individual carries at the SNP.

We first introduce a signed bipartite genotype and phenotype network (GPN) (**Figure 1.1a**). The weight of an edge represents the strength of the association between the two nodes (one is the phenotype and the other one is the genotype). The strength of the association has two directions, positive and negative. The adjacency matrix of GPN is a $K\times M$ matrix $\mathbf{T}=(T_{km})$, where T_{km} represents the strength of the association between the k^{th} phenotype and the m^{th} SNP. To calculate the adjacency matrix $\mathbf{T}$, we consider both the strengths and the directions of the associations. We first consider that there are no covariates. The strength of the association T_{km} can be estimated by the score test statistic $S_{km}=\sum_{i=1}^{n}(y_{ik}-\bar{y}_k)g_{im}$ and its p-value p_{km} under the generalized linear models $g\left(E\left(y_{ik}\,|\,g_{im}\right)\right)=\beta_{0km}+\beta_{1km}g_{im}$ ($k=1,\cdots,K$ and $m=1,\cdots,M$) [48]. Here, $\bar{y}_k=\sum_{i=1}^{n}y_{ik}/n$ and $g(\ )$ is a monotonic link function. Two commonly used link functions are the identity link for quantitative traits and the logit link for binary traits. If there are p covariates for the i^{th} individual, $x_{i1},\cdots,x_{ip}$, we adjust genotype and phenotype for the covariates using the following linear models proposed by Price et al.[49] and Sha et al.[50],

$$\begin{aligned} y_{ik}&=\alpha_{0k}+\alpha_{1k}x_{i1}+\cdots+\alpha_{pk}x_{ip}+\varepsilon_{ik}\\ g_{im}&=\gamma_0+\gamma_1x_{i1}+\cdots+\gamma_px_{ip}+\tau_{im}\end{aligned},$$

where $\boldsymbol{\varepsilon}_k=(\varepsilon_{1k},\cdots,\varepsilon_{nk})^T$ and $\boldsymbol{\tau}_m=(\tau_{1m},\cdots,\tau_{nm})^T$ denote the error terms of the k^{th} phenotype and the m^{th} SNP, respectively. We use the residuals of the respective linear model to replace the original genotypes and phenotypes.

For quantitative traits or binary traits with fairly balanced case-control ratios, we can use the normal approximation of $S_{km}\sim N(0,\sigma_{km}^2)$ to calculate p-value p_{km} under the null hypothesis that the k^{th} phenotype and the m^{th} SNP have no association, where $\sigma_{km}^2=\sum_{i=1}^{n}(y_{ik}-\bar{y}_k)^2\sum_{i=1}^{n}(g_{im}-\bar{g}_m)^2/n$ and $\bar{g}_m=\sum_{i=1}^{n}g_{im}/n$. Dey et al. [44] pointed out that a normal approximation of S_{km} has inflated type I error rates for binary traits with unbalanced case-control ratios. Therefore, we use saddlepoint approximation to calculate the p-value p_{km} for the phenotypes with unbalanced, especially extremely unbalanced

case-control ratios[44]. We define the $(k,m)^{th}$ element of the adjacency matrix of GPN, T_{km}, as $T_{km}=\text{sign}(S_{km})F_{Chi}^{-1}(1-p_{km})$, where $F_{Chi}(\ )$ denotes the CDF of χ_1^2. That is, we use $\text{sign}(S_{km})$ to define the direction of the association and use $F_{Chi}^{-1}(1-p_{km})$ to define the strength of the association. $T_{km}>0$ and $T_{km}<0$ represent two directions of the association between the k^{th} phenotype and the m^{th} SNP. If $T_{km}>0$, the minor allele of the m^{th} SNP is a protective allele to the k^{th} phenotype; if $T_{km}<0$, the minor allele of the m^{th} SNP is a risk allele to the k^{th} phenotype.

Although a bipartite network may give the most complete representation of a particular network, it is often convenient to work with just one type of nodes, that is, phenotypes or genotypes. The Phenotype and Phenotype Network (PPN) is the one-mode projection of GPN on phenotypes. In PPN, nodes only represent phenotypes (**Figure 1.1b**). Let $\mathbf{W}=(W_{kl})$ denote the adjacency matrix of the PPN in which each edge has a positive or negative weight. We define W_{kl} as the weight of the edge connecting the k^{th} and l^{th} phenotypes, which is given by

$$W_{kl}=\frac{\sum_{m=1}^{M}\left(T_{km}-\bar{T}_k\right)\left(T_{lm}-\bar{T}_l\right)}{\sqrt{\sum_{m=1}^{M}\left(T_{km}-\bar{T}_k\right)^2\sum_{m=1}^{M}\left(T_{lm}-\bar{T}_l\right)^2}},\quad k,l=1,\cdots,K.$$

Here, W_{kl} is the genetic correlation between the k^{th} and l^{th} phenotypes based on the association strengths T_{km} for $k=1,\cdots,K$ and $m=1,\cdots,M$. Thus, the PPN is also a signed network.

1.2.3 Community Detection Method

We apply a powerful community detection method to partition K phenotypes into disjoint network modules using the Ward hierarchical clustering method with a similarity matrix defined by the genetic correlation matrix $\mathbf{W}$[45]. The number of network modules is determined by the following perturbation procedure[51]. In details, we first use the Ward hierarchical clustering method to group the K phenotypes into k_0 ($k_0=1,\cdots,K-1$) clusters and build the $K\times K$ connectivity matrix $\mathbf{C}_{k_0}$ with the $(k,l)^{th}$ element of matrix $\mathbf{C}_{k_0}$ given by

$$\mathbf{C}_{k_0}(k,l)=\begin{cases}1, & \text{if phenotype } k \text{ and phenotype } l \text{ are in the same cluster}\\ 0, & \text{otherwise}\end{cases}.$$

Then, we generate B perturbed data sets. The b^{th} perturbed data set is generated by $T_{km}^{(b)}=T_{km}+\varepsilon_{km}$, where $\varepsilon_{km}\sim N(0,\sigma^2)$, $\sigma^2=\text{median}\left(\text{var}(\boldsymbol{T}_1),\cdots,\text{var}(\boldsymbol{T}_M)\right)$, and $\boldsymbol{T}_m=(T_{1m},\cdots,T_{Km})$. We denote the connectivity matrix of k_0 cluster based on the b^{th} perturbed data set by $\mathbf{C}_{k_0}^{(b)}$. Let $\mathbf{A}_{k_0}=\sum_{b=1}^{B}\mathbf{C}_{k_0}^{(b)}/B$ and $\mathbf{D}_{k_0}=\left|\mathbf{A}_{k_0}-\mathbf{C}_{k_0}\right|$, F_{k_0} denotes the

empirical CDF of the elements of $\mathbf{D}_{k_0}$, and AF_{k_0} denotes the area under the curve of F_{k_0}, where $F_{k_0}(x) = \#\left\{\mathbf{D}_{k_0}(l,k) \le x : l,k = 1,\cdots,K\right\} \Big/ K^2$. Then, the optimal number of network modules is given by

$$C = \arg\max_{k=1,\cdots,K-1} \left\{ \left| AF_{k+1} - AF_k \right| \right\}.$$

We can use the identified C network modules to further investigate the associations between phenotypes in each network module and SNPs.

1.2.4 Multiple Phenotype Association Tests

After we obtain C network modules for the phenotypes, we apply a multiple phenotype association test to identify the association between phenotypes in each of the C network modules and a SNP. Any multiple phenotype association test can be applied here. In this article, we apply six commonly used multiple phenotype association tests to each network module, including ceCLC[46], CLC[12], HCLC[47], MultiPhen[20], O'Brien[15], and Omnibus[12] (see details in **Text A.1**), then a Bonferroni correction is used to adjust for multiple testing for the C network modules to test if all phenotypes in the C network modules associated with a SNP.

1.2.5 Data Simulation

We conduct comprehensive simulation studies to evaluate the type I error rates and powers of multiple phenotype association tests based on network modules detected by GPN and compare them to the powers of the corresponding tests without considering network modules. To evaluate the performance of our proposed method, we consider different types of phenotypes: (i) mixture phenotypes: half quantitative and half qualitative with balanced case-control ratios, and (ii) binary phenotypes: all qualitative but with extremely unbalanced case-control ratios. We generate N individuals with M SNPs and K phenotypes. The genotypes at M SNPs are generated according to the minor allele frequency (MAF) under Hardy-Weinberg Equilibrium (HWE). Below, we first describe how to generate quantitative phenotypes. Suppose that there are C phenotypic categories and $k = K/C$ phenotypes in each phenotypic category. Let $\mathbf{Y}_c = (\boldsymbol{y}_{c1},\cdots,\boldsymbol{y}_{ck})$ denote the phenotypes in the c^{th} category. Similar to Sha et al. [12], we generate k quantitative phenotypes in each category using the following factor model,

$$\mathbf{Y}_c = \mathbf{G} \cdot \mathbf{B}_c + c_0 \cdot \boldsymbol{f}_c \cdot \boldsymbol{1}_k^T + \sqrt{1 - c_0^2} \cdot \mathbf{E}_c,$$

where $\mathbf{G} = (\boldsymbol{G}_1,\cdots,\boldsymbol{G}_M)$ is the matrix of M SNPs with dimension $N \times M$ which are generated from a $\text{binomial}(2, MAF)$ distribution for each SNP; $\mathbf{B}_c$ is an $M \times k$ matrix of effect sizes of M SNPs on k phenotypes in the c^{th} phenotypic category; $\mathbf{E}_c \sim MVN_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ is an $N \times k$ matrix of error term with $\boldsymbol{\Sigma} = (\sigma_{ij})$, where $\sigma_{ij} = \rho^{|i-j|}$ and ρ is a constant between 0 to 1; $\boldsymbol{f}_c$ is a factor vector in $\mathbf{f} = (\boldsymbol{f}_1,\cdots,\boldsymbol{f}_C)$ which follows

$MVN_C\left(\boldsymbol{0}, \boldsymbol{\Sigma}_f\right)$, where $\boldsymbol{\Sigma}_f = \left(1-\rho_f\right)\mathbf{I}_C + \rho_f \mathbf{J}_C$, $\rho_f = \text{corr}\left(\boldsymbol{f}_i, \boldsymbol{f}_j\right)$ if $i \neq j$, $\mathbf{J}_C$ is a $C \times C$ matrix with all elements of 1, and $\mathbf{I}_C$ is the identity matrix; c_0 is a constant number which represents a proportion. Therefore, the correlation between the i^{th} phenotype and the j^{th} phenotype within each category is $c_0^2 + \left(1-c_0^2\right)\rho^{|i-j|}$ and the between-category correlation is $c_0^2\rho_f$.

To generate a qualitative disease affection status, we use a liability threshold model based on a quantitative phenotype and its case-control ratio. Let n_a and n_c denote the number of affected individuals and the number of non-affected individuals. For a given case-control ratio r and sample size N, $n_c = N/\left(r+1\right)$ and $n_a = rN/\left(r+1\right)$. An individual is defined to be affected if the individual's phenotype is in the top n_a of all phenotypes. For each phenotype, the case-control ratio is randomly chosen from a set S. The set S contains all case-control ratios with the number of cases greater than 200 from UK Biobank ICD-10 code level 3 phenotypes (see Real Dataset).

Based on the factor model, we consider different numbe$C = 2$rs of phenotypes, 60, 80, and 100, and different sample sizes. For mixture phenotypes, the sample sizes are 2,000 and 4,000; for binary phenotypes, the sample sizes are 10,000 and 20,000. We consider the following six models (**Table A.1**) with $M = 2{,}000$, $MAF \sim U\left(0.05, 0.5\right)$, $\rho = 0.3$, $c_0^2 = 0.5$, and $\rho_f = 0.3/c_0^2$ (between-category correlation is 0.3). $\hat{\boldsymbol{\lambda}}_1 = \beta\left(1, \cdots, 1\right)^T$ and $\hat{\boldsymbol{\lambda}}_2 = \frac{2\beta}{k+1}\left(1, \cdots, k\right)^T$ are two types of effect sizes.

Model 1: $M_{\text{causal}} = 100$, $C = 2$, and all phenotypes are associated with at least one SNP with the same effect sizes but different directions. That is, the first 50 SNPs affect the phenotypes in the first category with $\hat{\boldsymbol{\lambda}}_1$ and the second 50 SNPs affect the phenotypes in the second category with $-\hat{\boldsymbol{\lambda}}_1$.

Model 2: $M_{\text{causal}} = 100$, , and all phenotypes are associated with at least one SNP with different effect sizes and different directions. That is, the first 50 SNPs affect the phenotypes in the first category with $\hat{\boldsymbol{\lambda}}_1$ and the second 50 SNPs impact the phenotypes in the second category with $-\hat{\boldsymbol{\lambda}}_2$.

Model 3: $M_{\text{causal}} = 100$, $C = 5$, and only phenotypes in the first two categories are associated with the first 100 SNPs with the same settings as in Model 1. The phenotypes in the remaining three categories do not associate with any SNPs.

Model 4: $M_{\text{causal}} = 100$, $C = 5$, and only phenotypes in the first two categories are associated with the first 100 SNPs with the same settings as in Model 2. The phenotypes in the remaining three categories do not associate with any SNPs.

Model 5: $M_{\text{causal}} = 200$, $C = 4$, and all phenotypes are associated with at least one SNP. That is, the first 50 SNPs affect the phenotypes in the first category with $\hat{\boldsymbol{\lambda}}_1$, the second 50 SNPs affect the phenotypes in the second category with $-\hat{\boldsymbol{\lambda}}_1$, the third 50 SNPs affect the phenotypes in the third category with $\hat{\boldsymbol{\lambda}}_2$, and the fourth 50 SNPs affect the phenotypes in the fourth category with $-\hat{\boldsymbol{\lambda}}_2$.

Model 6: $M_{\text{causal}} = 200$, $C = 10$, and only phenotypes in the first four categories are associated with the first 200 SNPs with the same settings as in Model 5. The phenotypes in the remaining six categories do not associate with any SNPs.

1.2.6 Comparison of Methods

We use six multiple phenotype association tests to evaluate the performance of our proposed method based on network modules. Therefore, we consider the following two types of comparisons.

Comparison 1: Apply six multiple phenotype tests without considering network modules.

We test the association between K phenotypes and a SNP. For each simulation model, we run B Monte-Carlo (MC) runs. The steps for the b^{th} MC run are as follow. i). Generate N individuals with M SNPs and K phenotypes in C categories; ii). Test the association between K phenotypes and M SNPs using each of the multiple phenotype association tests. The p-value for the m^{th} SNP in the b^{th} MC run is given by $p_m^{(b)}$. To evaluate the type I error rates of the tests, we generate phenotypes from the null model, that is, for each model, we set $\beta = 0$. The type I error rate, $\text{T1E}_{N.O.}$, can be calculated by

$$\text{T1E}_{N.O.} = \frac{\sum_{b=1}^{B}\sum_{m=1}^{M} I\left(p_m^{(b)} \le \alpha\right)}{B \times M}.$$

To evaluate power, we generate phenotypes from each of the six models with different effect sizes β. The power, $\text{power}_{N.O.}$, can be calculated by

$$\text{power}_{N.O.} = \frac{\sum_{b=1}^{B}\sum_{m=1}^{M_{\text{causal}}} I\left(p_m^{(b)} \le \alpha\right)}{B \times M_{\text{causal}}}.$$

Comparison 2: Apply six multiple phenotype tests by considering network modules.

For each simulation model, we run B MC runs. We use the following steps for b^{th} MC run. i). Generate N individuals with M SNPs and K phenotypes in C categories; ii). Construct the GPN based on the shared genetic architecture; iii). Detect $C^{(b)}$ network modules for the K phenotypes using the community detection method; iv). Test the association between phenotypes in each of the $C^{(b)}$ network modules and each of M SNPs using one of the six tests. We use $p_{cm}^{(b)}$ to denote the p-value of the associtio test between phenotypes in the c^{th} network module and the m^{th} SNP for $c = 1, 2, \cdots, C^{(b)}$. To evaluate

the type I error rate of a test based on the network modules, we generate phenotypes under the null model. That is, for each model, we set $\beta = 0$. The type I error rate, T1E_{NET}, can be calculated by

$$\text{T1E}_{NET} = \frac{\sum_{b=1}^{B}\sum_{m=1}^{M} I\left(\min_{c=1,\cdots,C^{(b)}}\left\{p_{cm}^{(b)}\right\} \le \alpha / C^{(b)}\right)}{B \times M}.$$

To evaluate power, we generate phenotypes for each model with different effect sizes β. The power, power_{NET}, can be calculated by

$$\text{power}_{NET} = \frac{\sum_{b=1}^{B}\sum_{m=1}^{M_{\text{causal}}} I\left(\min_{c=1,\cdots,C^{(b)}}\left\{p_{cm}^{(b)}\right\} \le \alpha / C^{(b)}\right)}{B \times M_{\text{causal}}}.$$

1.2.7 Real Dataset

The UK Biobank is a population-based cohort study with a wide variety of genetic and phenotypic information[52]. It includes ~ 500K people from all around the United Kingdom who were aged between 40 and 69 when recruited in 2006-2010[41,53]. Genotypes from the UK Biobank have extracted 488,377 participants with 784,256 variants in autosomal chromosomes. The preprocess of genotype is achieved by quality controls (QCs) which are performed on both SNPs and individuals using PLINK 1.9[54]. Same QCs as Liang et al.[47] (**Figure A.1**), we filter out SNPs with missing rates > 5%, Hardy-Weinberg equilibrium exact test p-values < 10^{-6}, and MAF < 5%. We also filter out individuals with missing rates > 5% and individuals without sex. After quality controls, 288,647 SNPs and 466,580 individuals remain for our next step analysis.

In this study, we define EHR-derived phenotypes using the ICD-10 codes, which is a standardized coding system for defining disease status as well as for billing purposes[6]. After truncating each full ICD-10 code to UK Biobank ICD-10 level 3 code, we consider 72 unique truncated ICD codes with the number of cases greater than 200 in Chapter XIII (Diseases of the musculoskeletal system and connective tissue), such as rheumatoid arthritis (M06.9), psoriatic arthropathies (M07.3), etc. Note that there are two phenotypes (M45: Ankylosing spondylitis and M45.X9: Ankylosing spondylitis (Site unspecified)) which are not truncated by the ICD-10 code digits, however, these two phenotypes are defined by UK Biobank level 3 code. For each individual, if a corresponding truncated ICD code ever appears, we denote the EHR-derived phenotype for that individual as "1", otherwise, we denote the EHR-derived phenotype for that individual as "0". After truncating ICD-10 codes, we generate a total of 502,591 individuals who have 72 EHR-derived phenotypes in Chapter XIII. Following the phenotype preprocess introduced in Liang et al.[47], 337,285 individuals are kept (**Figure A.1**).

After data preprocessing procedures, individuals with both genotype and phenotype information are used in our study. There is a complete set of 322,607 individuals across 288,647 SNPs with 72 EHR-derived phenotypes. Among the 72 phenotypes, lumbar and other intervertebral disk disorders with myelopathy (M51.0) has the smallest case-control ratio 0.000658 with 212 cases and 322,395 controls; Gonarthrosis (M17.9) has the largest

case-control ratio 0.03937 with 12,218 cases and 310,389 controls. Therefore, all of the phenotypes we considered in our analysis have extremely unbalanced case-control ratios. Furthermore, each phenotype is adjusted by 13 covariates, including age, sex, genotyping array, and the first 10 genetic principal components (PCs)[50]. The analysis is performed based on the adjusted phenotypes.

1.2.8 Correlation Analysis

To compare the genetic and phenotypic correlations among the 72 EHR-derived phenotypes, we apply cross-triat LDSC regression[27] to obtain the genetic correlation and phenotypic correlation which can provide useful etiological insights[27]. GWAS summary statistics are generated from the association between phenotype and genotype which are calculated by the saddlepoint approximation. We use the precomputed LD scores of European individuals in the 1000 Genomes project for high-quality HapMap3 SNPs ('eur_w_ld_chr'). For the phenotypic correlation, we consider 70 phenotypes excluding M79.6 (Enthesopathy of lower limb) and M67.8 (Other specified disorders of synovium and tendon), since the heritabilities of these two phenotypes estimated by LDSC are out of bounds. For the genetic correlation, we only consider 52 phenotypes exlcuding 20 phenotypes, where the heritabilities of these phenotypes are not significantly different from zero. We apply the K-means hierarchical clustering method to compare the correlations of phenotypes obtained by our proposed GPN and LDSC.

1.2.9 Post-GWAS Analyses

Pathway enrichment analysis. To better understand the biological functions behind the SNPs identified by one multiple phenotype association test, we identify the pathways in which the identified SNPs are involved. We use the functional annotation tool named Database for Annotation, Visualization, and Integrated Discovery bioinformatics resource (DAVID: https://david.ncifcrf.gov/)[55,56] for the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. A mapped gene used in the pathway enrichment analysis denotes the gene that includes at least one identified SNPs with a 20kb window region. The biological pathways with FDR < 0.05 and enriched gene count > 2 are considered statistically significant[57].

Tissue enrichment analysis. To prioritize and interpret the GWAS signals and identify lead SNPs, tissue enrichment analyses are performed using the Functional Mapping and Annotation (FUMA: https://fuma.ctglab.nl/)[58] platform and the GWAS signals from one multiple phenotype association test in N.O. and in NET, respectively. FUMA first performs a genic aggregation analysis of GWAS association signals to calculate gene-wise association signals using MAGMA, which is a commonly used generalized gene-set analysis of GWAS summary statistics[59]. Then, it subsequently tests whether tissues and cell types are enriched for expression of the genes with gene-wise association signals. For tissue enrichment analysis, we use 30 general tissue types in GTEx v8 reference set (https://gtexportal.org/home/).

Colocalization analysis. As most associated variants are noncoding, it is expected that they influence disease risk through altering gene expression or splicing[60]. The colocalization analysis is a way to identify the association of a GWAS SNP and a gene expression QTL that are colocalized. We perform colocalization analysis using the 'coloc' package in R[61],

a Bayesian statistical methodology that tests pairwise colocalization of eQTLs with unique identified SNPs by ceCLC in NET and N.O. from the UK Biobank dataset. The SNP-gene associations in the Muscle Skeletal tissue are downloaded from GTEx v7. We use the default of prior probabilities, $p_1 = p_2 = 10^{-4}$ and $p_{12} = 10^{-5}$, for a causal variant in an eQTL or a GWAS SNP and a shared causal variant between eQTL and GWAS SNP, respectively.

1.3 Results

1.3.1 Simulation studies

We first use extensive simulation studies to validate multiple phenotype association studies based on the newly proposed GPN. In the simulation studies, we assess the type I error rate and power with different numbers of phenotypes (60, 80, and 100), different types of phenotypes along with different sample sizes: (i) mixture phenotypes are half quantitative and half qualitative with balanced case-control ratios for sample sizes of 2,000 and 4,000, and (ii) binary phenotypes are all qualitative but with extremely unbalanced case-control ratios for sample sizes of 10,000 and 20,000. Similar to the simulation models introduced in Sha et al.[12], we generate six different models (see Data Simulation for a full description of the simulation models).

Type I Error Rates.

Table A.2-A.7 summarize the estimated type I error rates of six multiple phenotype association tests for mixture phenotypes under models 1-6, respectively. "N.O." represents the type I error rates of multiple phenotype association tests being calculated without considering network modules; "NET" presents the type I error rates of the tests being evaluated by considering network modules detected by GPN. Based on 500 Monte-Carlo (MC) runs which is the same as 10^6 replicates, the 95% confidence intervals (CIs) for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804,1.196), respectively. The bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. Almost all of the estimated type I error rates of ceCLC, CLC, HCLC, and Omnibus tests are within 95% CIs. However, O'Brien in NET has inflated type I error rates under model 6. MultiPhen has inflated type I error rates for the sample size of 2,000. If the sample size is 4000, MultiPhen in N.O. also inflates type I error rates, but MultiPhen in NET can control type I error rates for the significance level is 0.0001. **Table A.8-A.13** summarize the estimated type I error rates of six tests for binary phenotypes with extremely unbalanced case-control ratios under models 1-6. Similar to **Tables A.2-A.7**, ceCLC, CLC, HCLC, and Omnibus have corrected type I error rates at almost all simulation settings. However, O'Brien in NET has inflated type I error rates and MultiPhen has inflated type I error rates at all scenarios.

Power comparisons.

For power comparisons, we consider 100 causal SNPs for models 1-4 and 200 causal SNPs for models 5-6 (see Data Simulation). In each of the simulation models, the power is evaluated using 10 MC runs which is the same as 1,000 replicates for models 1-4 and 2,000 replicates for models 5-6. Meanwhile, the power is evaluated at the Bonferroni corrected significance level of 0.05 based on the number of causal SNPs in each MC run.

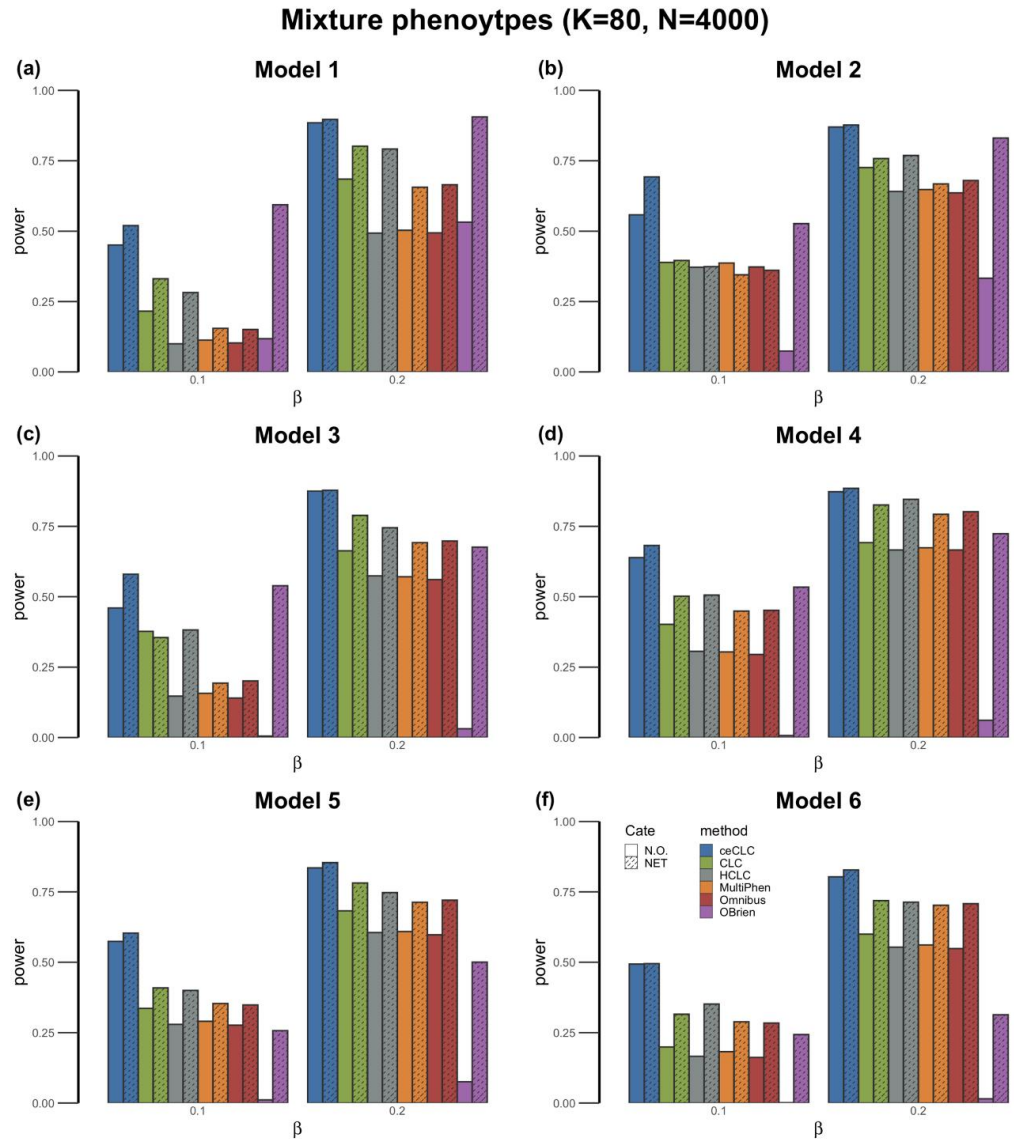


Figure 1.2. Power comparisons of the six tests as a function of effect size β under the six models. The number of mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) is 80 and the sample size is 4,000. The power of all of the six tests is evaluated using 10 MC runs.

Figure 1.2 (**Figure A.2**) shows the power of six multiple phenotype association tests under six simulation models for different effect sizes with a total of 80 mixture phenotypes and a sample size of 4,000 (2,000). From **Figure 1.2** and **Figure A.2**, we can see that: (i) All tests in NET (filled by the dashed line) are much more powerful than those in N.O., indicating that tests based on network modules detected by GPN are more powerful than the tests without considering network modules. Since the community detection method can partition phenotypes into different network modules based on shared

genetic architecture, the phenotypes can be clustered in the same module if they have higher genetic correlations. In particular, the power of O'Brien[15] increases a lot in the case of a SNP affecting phenotypes in different directions. (ii) ceCLC is more powerful than other tests in both N.O. and NET under the six simulation models. (iii) As sample size increases, the power of all multiple phenotype association tests increases. We also perform power comparisons for a total of 60 and 100 mixture phenotypes with 2,000 and 4,000 sample sizes for different effect sizes under the six simulation models (**Figures A.3- A.6**), respectively. We observe that the patterns of the power are similar to those observed in **Figure 1.2** and **Figure A.2**.

To mimic phenotypes in the UK Biobank, we also consider the case with all phenotypes being binary with extremely unbalanced case-control ratios. The phenotypes are generated based on extremely unbalanced case-control ratios which are randomly selected from the set of case-control ratios with cases greater than 200 from UK Biobank ICD-10 code level 3 phenotypes (case-control ratios belong to $[0.000658, 0.03937]$). In this simulation, we consider a total of 60, 80, and 100 phenotypes along with two sample sizes, 10,000 and 20,000. **Figures A.7-A.12** show the power comparisons of the six tests under six simulation models. The patterns of power comparisons for binary phenotypes are similar to those observed in **Figure 1.2** and **Figure A.2-A.6**.

1.3.2 Real Data Analysis based on UK Biobank

Furthermore, we apply the newly proposed multiple phenotype association test based on network modules detected by GPN to a set of diseases of the musculoskeletal system and connective tissue across more than 300,000 individuals from the UK Biobank.

Network Module Detection.

We construct GPN based on 72 EHR-derived phenotypes in the diseases of the musculoskeletal system and connective tissue with 288,647 SNPs in autosomal chromosomes in the UK Biobank. Due to all phenotypes in our analysis being extremely unbalanced, the strength of the association between phenotype and genotype is calculated by the saddlepoint approximation[44]. After the construction of GPN, we apply a powerful community detection method and these 72 phenotypes are partitioned into 8 disjoint network modules (**Figure 1.3**). There are 2-37 phenotypes in each module.

We can see that the network modules are not consistent with the ICD-based categories which are based on the underlying cause of death rather than the shared genetic architecture among all complex diseases. For example, **Figure 1.3** shows three phenotypes, M32.9 Systemic lupus erythematosus, M35.0 Sicca syndrome, and M65.3 Trigger finger, are detected in network module III (in red). However, these three phenotypes do not belong to the same ICD-category (Data-Field 41202 in UK Biobank), where M35.0 is one of the diseases in the other systemic involvement of connective tissue (M35) and M65.3 belongs to the synovitis and tenosynovitis (M65). To investigate the genetic correlation among these three phenotypes, we use the saddlepoint approximation to test the association between each phenotype and each SNP. As shown in **Figure A.13**, the Manhattan plots for the three phenotypes in network module III (M32.9, M35.0, and M65.3) have a similar pattern. Although the synovitis and tenosynovitis (M65.9) and M65.3 belong to the same

ICD code category (M65), the Manhattan plot of M65.9 shows that there are no SNPs significantly associated with this phenotype and the genetic correlation between M65.9 and M65.3 is not strong. Therefore, we can conclude that the community detection method based on our proposed GPN can partition phenotypes into different categories based on the shared genetic architecture.

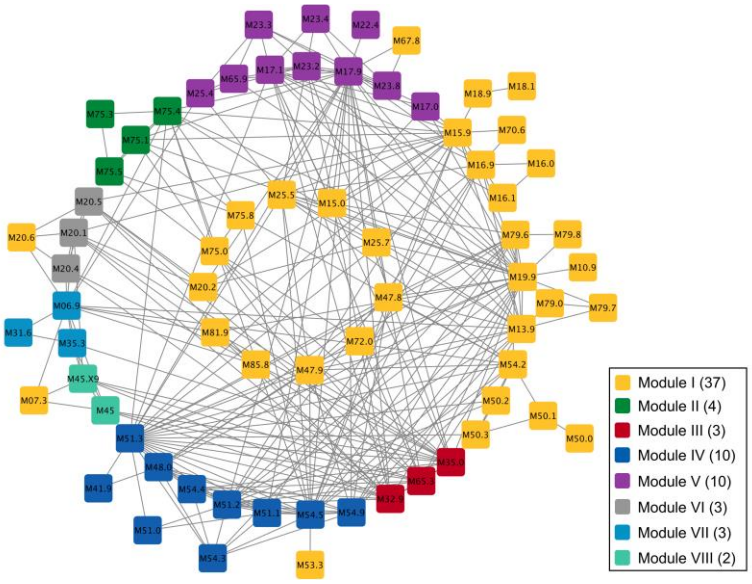


Figure 1.3. The network modules detected by the powerful community detection method based on GPN. The blocks with different color indicate different modules, where the values in the legend represent the number of phenotypes in each network module. The labels of phenotypes are listed in the form of ICD-10 code and the corresponding diseases can be found in the UK Biobank. The connection between two phenotypes represents the absolutely value of the weight greater than 40. The graph was prepared by Cytoscape.

Furthermore, we apply the hierarchical clustering method to compare the genetic correlation of phenotypes obtained by our proposed GPN and that estimated by LDSC [27]. **Figures A.14-A.15** show that dendrograms of hierarchical clustering method based on the genetic correlation of phenotypes obtained by GPN, and the phenotypic or genetic correlation estimated by LDSC, respectively. In **Figure A.14**, the cluster results of the phenotypic correlation estimated by LDSC are similar to that of the genetic correlation based on GPN, but GPN can separately identify two highly genetic correlated phenotypes, ankylosing spondylitis (M45) and ankylosing spondylitis with site unspecified (M45.X9). However, the cluster results of the genetic correlation estimated by LDSC are different from those obtained by GPN. Some phenotypes in the same UK Biobank level 1 category can be clustered in the same group by GPN but not by LDSC (**Figure A.15**).

Interpretation of the Association Test.

We apply five multiple phenotype tests (ceCLC, CLC, HCLC, O'Brien, and Omnibus) to test the association between 72 EHR-derived phenotypes and each of 288,647 SNPs in the UK Biobank. MultiPhen is not considered here since it has inflated type I error rates, especially for the phenotypes with extremely unbalanced case-control ratios.

First, we apply the five tests in N.O. to test the association between 72 phenotypes and each SNP. We use the commonly used genome-wide significance level 5×10^{-8}. **Figure 1.4(a)** shows the Venn diagram of the number of SNPs identified by the five tests. There are 11 SNPs identified by all five tests. ceCLC identifies 647 SNPs with 32 unique SNPs not being identified by other four tests. Among the 32 novel SNPs, two SNPs, rs13107325 (p-value = 4.6×10^{-10}) and rs443198 (p-value = 1.73×10^{-11}), are significantly associated with at least one of the 72 phenotypes reported in the GWAS catalog (**Table A.14**). rs13107325 is reported to be associated with osteoarthritis (M19.9) [62] and rotator cuff syndrome (M75.1) [63]. Meanwhile, rs13107325 is mapped to gene *SLC39A8* that is also reported to be significantly associated with multisite chronic pain (M25.5) [64]. rs443198 is mapped to gene *NOTCH4* which is associated with systemic sclerosis (M34) [65]. Moreover, the mapped gene *NOTCH4* is one of the most important genes reported to be associated with multiple diseases in the disease category of the musculoskeletal system and connective tissue, such as rheumatoid arthritis (M06.9) [66], psoriatic arthritis (M07.3) [67], Takayasu arteritis (M31.4) [68], systemic lupus erythematosus (M32.9) [69], and appendicular lean mass (M62.9) [70]. We map these 32 unique SNPs into genes with 20 kb upstream and 20 kb downstream regions. There are 27 out of 32 SNPs with corresponding mapped genes associated with 14 phenotypes reported in the GWAS catalog (**Table A.14**). These 14 phenotypes and corresponding ICD-10 codes are summarized in **Table A.15**.

Next, we test the associations between phenotypes in each of the eight network modules detected by the GPN and each SNP. Then, we adjust the p-value of each method for testing the association between a SNP and all of the 72 phenotypes by Bonferroni correction. We adopt the commonly used genome-wide significance level 5×10^{-8}. **Figure 1.4(b)** shows that all tests can identify more SNPs comparing with the number of SNPs identified in N.O. ceCLC in NET identifies 980 SNPs, where 647 SNPs are identified in N.O. Meanwhile, there are 950 SNPs identified by HCLC, 949 SNPs by CLC, and 891 SNPs by Omnibus, where the corresponding results in N.O. are 354 SNPs, 808 SNPs, and 634 SNPs, respectively. In particular, the number of SNPs identified by O'Brien in NET is increased a lot, where there are 948 SNPs identified in NET and only 57 SNPs identified in N.O. As the results shown in **Figure 1.4(b)**, there are 807 overlapped SNPs identified by all five tests in NET which is much larger than 11 overlapped SNPs identified in N.O.

To compare the difference between the tests in N.O. and in NET, we summarize the number of overlapping SNPs identified by each method in N.O. and NET in **Figure A.17**. We observe that most SNPs identified in N.O. can be identified in NET. Meanwhile, tests in NET can identify much more SNPs than those in N.O. As mentioned previously, the advantage of the tests based on the network modules detected by GPN is that we can identify potential pleiotropic SNPs and also interpret SNP effects on which network modules based on the shared genetic architecture. Notably, we also investigate the smallest p-value obtained by each of the eight phenotypic modules for each of the 980 SNPs identified by ceCLC. For example, 396 SNPs have the smallest p-values for testing the association with network module III. Based on the results of the univariate score test corrected for saddlepoint approximation (SPAtest) (**Figure A.13**), 104 SNPs are significantly associated with at least one phenotype in module III. All of these 104 SNPs can be identified by ceCLC, HCLC, and Omnibus and 103 SNPs can be identified by CLC

and O'Brien in NET. The results show that the tests based on network modules can detect potential pleiotropic loci which can not be detected by the univariate test.

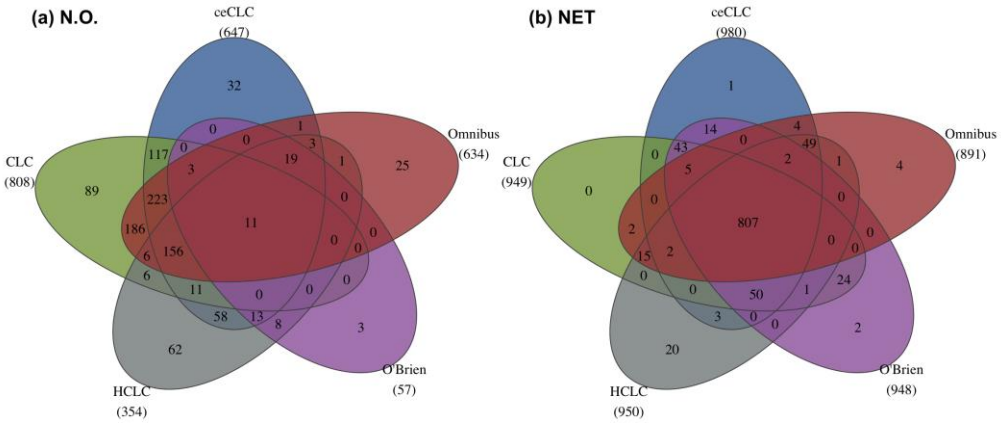


Figure 1.4. The Venn diagram of the number of SNPs identified by ceCLC, CLC,HCLC, O'Brien, and Omnibus in N.O. (**a**) and in NET (**b**). The number below each method indicates the total number of SNPs identified by the corresponding method.

Pathway Enrichment Analysis.

ceCLC is more powerful than the other four tests in simulations and also can identify more SNPs in real data analysis, therefore, we only perform the post-GWAS analyses of the SNPs identified by ceCLC. There are 191 mapped genes containing at least one of the 647 SNPs identified by ceCLC in N.O. and 252 mapped genes containing at least one of the 980 SNPs identified by ceCLC in NET. In this study, significantly enriched pathways are identified by those genes with false discovery rate (FDR) < 0.05.

From the pathway enrichment analyses, we observe that ceCLC based on the network modules identifies more significantly enriched pathways than that without considering network modules. **Figure 1.5** shows that 16 pathways are significantly enriched by 191 mapped genes in N.O. and 29 pathways are significantly enriched by 252 mapped genes in NET, where all of the 16 pathways identified in N.O. are also identified in NET. Two pathways identified in N.O. and NET, rheumatoid arthritis (hsa05323; FDR = 8.72×10^{-3} in N.O. and FDR = 6.48×10^{-8} in NET) and systemic lupus erythematosus (hsa05322; FDR = 4.25×10^{-19} in N.O. and FDR = 1.02×10^{-40} in NET) showed in **Figure 1.5**, are related to the diseases of the musculoskeletal system and connective tissue. For example, osteopetrosis (M19.9) and rheumatoid arthritis (M06.9) are related to the rheumatoid arthritis pathway. Meanwhile, the pathway related to at least one of the 72 phenotypes, hematopoietic cell lineage (hsa04640; FDR = 1.08×10^{-5}), is only identified in NET. Notably, DBGET (https://www.genome.jp/dbget-bin/www_bget?hsa05322) reports that there are two pathways related to systemic lupus erythematosus: antigen processing and presentation (hsa04612; FDR = 4.83×10^{-3} in N.O. and FDR = 2.82×10^{-16} in NET) identified in both N.O. and NET and cell adhesion molecule (hsa04514; FDR = 1.04×10^{-5}) only identified in NET.

Meanwhile, the above five pathways related to the diseases of the musculoskeletal system and connective tissue contain more enriched genes identified by ceCLC in NET than the enriched genes identified in N.O. For example, 43 SNPs within six mapped genes identified by ceCLC in N.O. are enriched in rheumatoid arthritis pathway, including *ATP6V1G2*, *HLA-DRA*, *LTB*, *TNF*, *HLA-DRB1*, and *HLA-DQA1*; and 111 SNPs within 12 mapped genes in NET are enriched in this pathway, including *HLA-DMA*, *HLA-DMB*, *ATP6V1G2*, *HLA-DRA*, *LTB*, *HLA-DOA*, *TNF*, *HLA-DOB*, *HLA-DQA2*, *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1*. Compared with the results of ceCLC in N.O., the test based on network modules identifies six more enriched genes, especially, gene *HLA-DMB* (including rs241458; p-value = 7.09×10^{-9}) and gene *HLA-DOA* (including rs3097646; p-value = 5.50×10^{-9}) that have not been reported in the GWAS catalog.

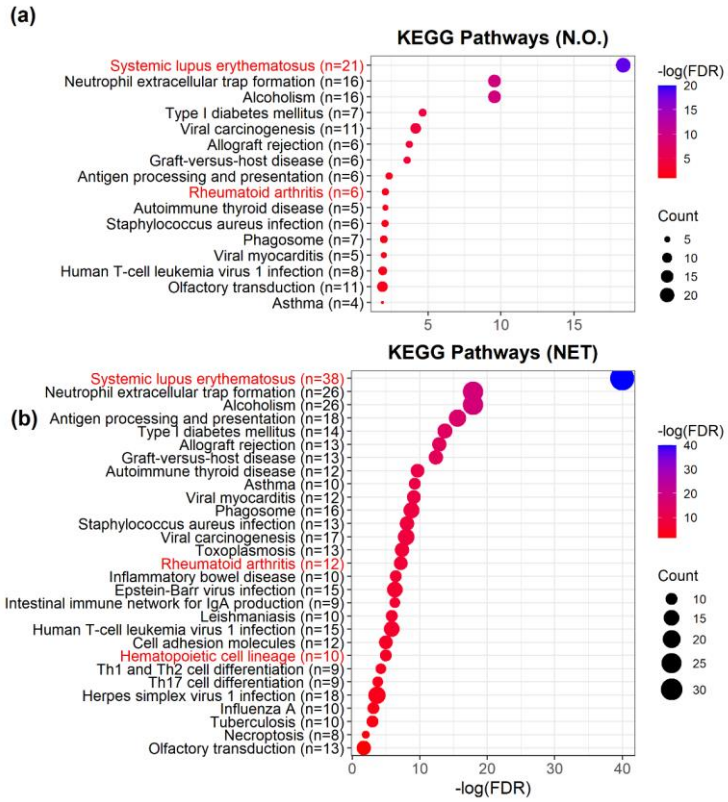


Figure 1.5. The results for the pathway enrichment analysis based on the genes identified by ceCLC and the KEGG database in N.O. (**a**) and NET (**b**). The red marked pathways denote the pathways related to the diseases of the musculoskeletal system and connective tissue. There are 191 genes in N.O. and 252 genes in NET that are applied to the pathway enrichment analysis.

Tissue Enrichment Analysis.

To further investigate the biological mechanism, we use FUMA[58] to annotate 191 mapped genes in N.O. and 252 mapped genes in NET in terms of biological context. Due to these mapped genes associated with at least one phenotype in the diseases of the musculoskeletal system and connective tissue, we can test if these mapped genes are enriched in the relevant-tissue based on FUMA. **Figure A.17** shows the ordered enriched tissues based on

the mapped genes identified by ceCLC in N.O. and NET. We observe that the mapped genes identified by ceCLC in N.O. are most enriched in brain-related tissue (**Figure A.17(a)**). Nevertheless, **Figure A.17(b)** shows that the mapped genes identified by ceCLC in NET are significantly enriched in the Muscle-Skeletal tissue with p-value < 0.05. The construction of GPN is benefit to multiple phenotype association studies by clustering the related phenotypes based on the genetic information. Notably, the identified SNPs are more likely to be within the same relevant biological context.

Colocalization of GWAS and eQTL analysis.

We perform the colocalization analysis on the 33 unique SNPs identified by ceCLC (**Table A.14**; one SNP in NET and 32 SNPs in N.O.) and all SNP-gene association pairs in the Muscle Skeletal tissue reported in GTEx. **Figure A.18** shows the colocalization signals with the uniquely identified SNPs by ceCLC that are selected to be the lead SNPs in the colocalization analysis. NET identifies one unique SNP, rs4148866, which is mapped to gene *ABCB9*. Even if gene *ABCB9* has no reported associations with any diseases of the musculoskeletal system and connective tissue in the GWAS Catalog, the Bayesian posterior probability of colocalization analysis for shared variant of significant SNPs identified by ceCLC and gene expression in the Muscle Skeletal tissue (PP_{H4}) is 98.4%. The higher value of PP_{H4} indicates that gene *ABCB9* and Muscle Skeletal tissue play an important role in the disease mechanism due to the same variant responsible for a GWAS locus and also affecting gene expression[61]. Among 32 unique SNPs identified by ceCLC in N.O., there are two SNPs, rs34333163 and rs6916921, selected to be the lead SNPs (**Figure A.18**). Both of them are reported in the GWAS Catalog that have associations with at least one of the diseases in the musculoskeletal system. However, the PP_{H4} values for the corresponding genes *SLC38A8* and *ATP6V1G2* are lower than 50%.

1.4 Discussion

In this paper, we propose a novel method for multiple phenotype association studies based on genotype and phenotype network. The construction of a bipartite signed network, GPN, is to link genotypes with phenotypes using the evidence of associations. To understand pleiotropy in diseases and complex traits and explore the genetic correlation among phenotypes, we project genotypes into phenotypes based on the GPN. We also apply a powerful community detection method to detect the network modules based on the shared genetic architecture. In contrast to previous community detection methods for disease networks, the applied method benefits from exploring the biological functionality interactions of diseases based on the signed network. Furthermore, we apply several multiple phenotype association tests to test the association between phenotypes in each network module and a SNP. Extensive simulation studies show that all multiple phenotype association tests based on network modules have corrected type I error rates if the corresponding test is a valid test for testing the association between a SNP and phenotypes without considering network modules. Most tests in NET are much more powerful than those in N.O. Meanwhile, we evaluate the performance of the association tests based on network modules detected by GPN through a set of 72 EHR-derived phenotypes in the diseases of the musculoskeletal system and connective tissue across more than 300,000 samples from the UK Biobank. Compared with the tests in N.O., all tests based on network

modules detected by GPN can identify more potentially pleiotropic SNPs and ceCLC can identify more SNPs than other methods.

In addition, the construction of GPN does not require access to individual-level genotypes and phenotypes data, which only requires association evidence between each genotype and each phenotype. Therefore, when individual-level data are not available, this evidence can be obtained from GWAS summary statistics, such as the effect sizes (odds ratios for binary phenotypes) and corresponding p-values. Meanwhile, the simulation studies show that the powerful network community detection method can correctly partition phenotypes into several disjoint network modules based on the shared genetic architecture. Since the determination of the number of network modules by applying community detection method is independent of the association tests45, we only need to perform the perturbation procedure once in real data analyses. In our real data analysis with 72 phenotypes and 288,647 SNPs, it only takes 1.5 hour with 1,000 perturbations to obtain the optimal number of network modules on a macOS (2.7 GHz Quad-Core Intel Core i7, 16 GB memory).

In summary, the proposed GPN provides a new insight to investigate the genetic correlation among phenotypes. Especially when the phenotypes have extremely unbalanced case-control ratios, the weight of an edge in the signed bipartite network can be calculated based on the saddlepoint approximation. The power of multiple phenotype association tests based on network modules detected by GPN are improved by incorporating the genetic information into the phenotypic clustering. Therefore, the proposed method can be applied to large-scale data across multiple related traits and diseases (i.e., biobanks data set, etc.).

1.5 Availability of data and materials

Data

The UK Biobank data are accessed via https://www.ukbiobank.ac.uk/[41].

The GWAS catalog summary data are accessed via https://www.ebi.ac.uk/gwas/.

The SNP-gene associations in the Muscle Skeletal tissue are downloaded via https://gtexportal.org/home/.

Software

The software for the proposed method is publicly available at https://github.com/xueweic/GPN.

PLINK version 1.9 can be downloaded from https://www.cog-genomics.org/plink/1.9/ [54].

LDSC: the command line tool for estimateing heritability and genetic correlation from GWAS summary statistiscs can be downloaded from https://github.com/bulik/ldsc[27].

FUMA: the platform that can be used to annotate, prioritize, visualize and interpret GWAS results can be found from https://fuma.ctglab.nl/[58].

DAVID: the functional tool can be found from https://david.ncifcrf.gov/[55,56].

Cytoscape: an open source software platform for visualizing complex networks which can be accessed via https://cytoscape.org/[71].

2 Chapter 2

Constructing genotype and phenotype network helps reveal disease heritability and phenome-wide association studies

Abstract

Analyses of a bipartite Genotype and Phenotype Network (GPN), linking the genetic variants and phenotypes based on statistical associations, provide an integrative approach to elucidate the complexities of genetic relationships across diseases and identify pleiotropic loci. We first assess contributions to constructing a well-defined GPN with a clear representation of genetic associations by comparing the network properties with a random network, including connectivity, centrality, and community structure. Next, we construct network topology annotations of genetic variants that quantify the possibility of pleiotropy and apply stratified linkage disequilibrium (LD) score regression to 12 highly genetically correlated phenotypes to identify enriched annotations. The constructed network topology annotations are informative for disease heritability after conditioning on a broad set of functional annotations from the baseline-LD model. Finally, we extend our discussion to include an application of bipartite GPN in phenome-wide association studies (PheWAS). The community detection method can be used to obtain a priori grouping of phenotypes detected from GPN based on the shared genetic architecture, then jointly test multiple phenotypes in each network module and one genetic variant to discover the cross-phenotype associations and pleiotropy. Significance thresholds for PheWAS are adjusted for multiple testing by applying the false discovery rate (FDR) control approach. Extensive simulation studies and analyses of 633 EHR-derived phenotypes in the UK Biobank GWAS summary dataset reveal that most multiple phenotype tests based on GPN can well-control FDR and identify more significant genetic variants comparing with the tests based on UK Biobank categories.

Keywords: genotype and phenotype network, network topology annotation, disease heritability, phenome-wide association studies, GWAS summary statistics

2.1 Introduction

The studies based on the biological networks have proven to be successful in providing a comprehensive understanding of the complex relationships that exist within the biological systems, such as gene regulatory networks[72,73], protein-protein interaction networks[74], human disease networks[30], et al. In particular, the human disease network is usually describing the system as a bipartite network, explicitly including two different types of nodes, in which diseases are connected to their associated genes. Rather than simply identifying the association between a genetic variant and a specific disease, constructing a bipartite network is presented the opportunity to explore the integrated molecular underpinnings of diseases[75]. Therefore, it can be used to explore whether human diseases

or complex traits and the corresponding genetic variants are related to each other at a higher level of cellular and organization[76,77]. In addition, due to many complex diseases being affected by a shared set of pleiotropic variants, the construction of a bipartite network can also be used to determine the pathobiological relationship of one disease to other diseases[75] and elucidate the complexities of genetic relationships across diseases[76].

Over the past decade, genome-wide association studies (GWAS) have generated an impressive list of genetic variant and phenotype association pairs[3,78], which offer a great opportunity to establish a bipartite network connecting genetic variants and phenotypes, referred to as a genotype and phenotype network (GPN)[77]. GPN provides integrative analyses to characterize complex relationships between genetic variants and phenotypes that are reproducible and accurately represent biological relationships and is thus of increasingly significant importance[31,79]. Notably, the construction of a well-defined GPN is crucial as it comes up with a clear representation of the genetic association between genetic variants and phenotypes, including connectivity, centrality, community structure, et al. Meanwhile, the real-world biological network, including GPN, often exhibits a scale-free degree distribution[80,81], which means that a small number of nodes (genetic variants and phenotypes) have a much larger number of connections than the majority of nodes. In a random network, the nodes are connected randomly without any preferential attachment, resulting in a network with a relatively uniform degree distribution[82]. Therefore, comparing the degree distribution of a bipartite GPN to that of a random network can reveal important insights into the underlying mechanisms driving the construction of the network. Additionally, random networks can serve as a useful null model for testing the significance of network properties observed in the bipartite GPN.

The centralities of the bipartite GPN are one of the most important properties that measure the importance of genetic variants (phenotypes) across phenotypes (genetic variants) based on the connectivity in the network[33]. The nodes with high centrality often act as hubs for information flow within the network[83]. For example, a genetic variant with high centrality accounting for all phenotypes is more likely to be a pleiotropic variant, as it is highly connected to multiple phenotypes in a bipartite GPN. Therefore, these centralities can be used to define the network topology annotations of genetic variants that quantify the possibility of a genetic variant being a pleiotropic variant. To study whether these network topology annotations are enriched for disease heritability, we apply stratified linkage disequilibrium (LD) score regression (S-LDSC)[84,85] along with the leave-one-phenotype-out strategy to quantify the contribution of these annotations to disease heritability. We condition our analyses of the network topology annotations on the baseline-LD model, which includes a broad set of coding, conserved, regulatory, and LD-related functional annotations[86]. Meanwhile, in a bipartite GPN, a phenotype with a higher centrality accounting for all genetic variants is more likely to have a higher heritability, as it is connected to multiple genetic variants or with higher association evidence.

With the widespread availability of electronic health records (EHR) data, phenome-wide association studies (PheWAS) have been used to systematically examine the impact of one genetic variant across a broad range of phenotypes. Phenotypes in the whole phenome can be grouped by digitized codes (e.g., ICD-10 code) to represent the common clinical factors underlying the diseases. However, the taxonomy of digitized codes depends

on their etiology rather than their genetic architecture. As a consequence, applying the community detection method for GPN allows us to identify network modules that provide an integrative approach to understanding the complex genetic relationships across phenotypes[77]. A network module is loosely defined as a subnetwork with high local link density so that the phenotypes within a network module share more genetic architecture across all genetic variants than phenotypes outside the network module[87,88]. Therefore, the network modules can serve as an a priori grouping of phenotypes in PheWAS, then we can jointly test multiple phenotypes in each network module and a genetic variant to discover the cross-phenotype associations and pleiotropy. For multiple testing corrections, we apply a refined false discovery rate (FDR) control approach to obtain the significance thresholds for PheWAS.

2.2 Methods and Materials

In this section, we first describe our approach to construct Genotype and Phenotype Networks (GPN; section 2.2.1) and define the network topology annotations for genetic variants and phenotypes (section 2.2.2). The construction of GPN does not require access to the individual-level genotypes and phenotypes data and only requires the marginal association evidence between each genetic variant and each phenotype (e.g., z-scores or estimated effect sizes from GWAS summary datasets). We identify differences in both denser representation and sparse representations of GPN with various sparsity approaches. We then provide details of the implementation of these approaches, such as heritability enrichment of network topology annotations (section 2.2.3), estimation of the genetic correlation of multiple phenotypes and community detection of phenotypes (section 2.2.4), and phenome-wide association studies (section 2.2.5). **Figure 2.1** shows the overview of this study.

2.2.1 Bipartite genotype and phenotype networks construction

We consider GWAS summary statistical results from the same or different study cohorts with K phenotypic traits. Assume that the GWAS summary results for the k^{th} ($k=1,\cdots,K$) phenotype are calculated by testing the marginal association between a genetic variant and the k^{th} phenotype based on a sample with N_k unrelated individuals. Note that $N_k = N_l$ ($k \neq l$) if the GWAS summary data of the k^{th} phenotype and l^{th} phenotype are calculated from the same study cohort, otherwise, $N_k \neq N_l$. For simplicity, we assume the generalized linear regression, $g\left(E\left(y_{ik} \mid g_{im}\right)\right) = \alpha_{0mk} + \boldsymbol{\alpha}_{mk}^T \boldsymbol{X}_{ik} + \beta_{mk} g_{im}$, where y_{ik} is the k^{th} phenotype value and $\boldsymbol{X}_{ik}$ is the vector of covariates, for example, used to account for population stratification in the study, for the i^{th} ($1 \le i \le N_k$) individual and the k^{th} phenotype. Assuming that there are M_k genetic variants in the GWAS summary statistics for the k^{th} phenotype and g_{im} is the genotype of the m^{th} ($1 \le m \le M_k$) genetic variant taking values from 0, 1, and 2 that counts the number of copies of the minor allele. Here, $g(\bullet)$ is either the identity link function for quantitative phenotypes or the logit link for binary phenotypes.

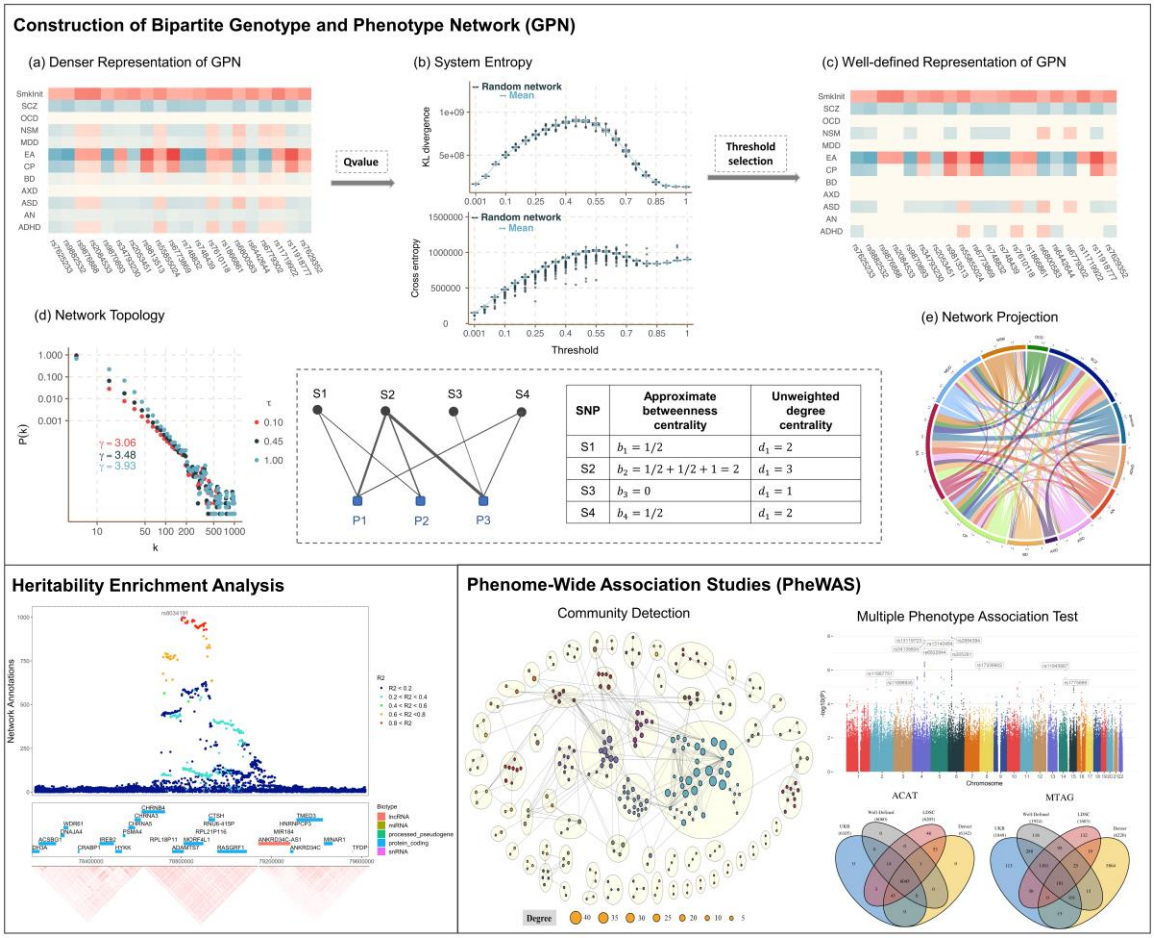


Figure 2.1. Graphical Abstract. Construction of bipartite genotype and phenotype network (GPN) includes: (a) – (c) Construction of the denser and well-defined representations of GPN by comparing the network properties with the random networks, including connectivity, centrality, and system entropy; (d) The weighted degree distributions with different thresholds and the examples of two network topology annotations, approximate betweenness centrality and degree centrality, used in the heritability enrichment analysis; (e) The one-mode projection of GPN onto phenotypes that are linked through shared genetic architecture. Heritability enrichment analysis and phenome-wide association studies are introduced as two important applications of the constructed GPN.

The GWAS summary results are calculated for testing the genetic association between the k^{th} phenotype and the m^{th} genetic variant under the null hypothesis $H_{0,mk}:\beta_{mk}=0$. The commonly used Wald-type statistic is defined as $Z_{mk}=\hat{\beta}_{mk}\big/se\left(\hat{\beta}_{mk}\right)$ under the generalized linear regression model, where $\hat{\beta}_{mk}$ is the maximum likelihood estimation (MLE) of β_{mk} and $se\left(\hat{\beta}_{mk}\right)$ is its estimated standard error[89]. The p-value p_{mk} may also be calculated by assuming $Z_{mk}\sim N(0,1)$ in the GWAS summary results.

Let M be the total number of unique SNPs included in the GWAS summary statistics for K phenotypes with the property of $\max_{k=1,\cdots,K}\{M_k\} \le M \le \sum_{k=1}^{K} M_k$. In particular, $M = \max_{k=1,\cdots,K}\{M_k\}$ if and only if there is at least one GWAS summary data containing all unique genetic variants and $M = \sum_{k=1}^{K} M_k$ if and only if there are no variants included in different GWAS summary data. We can exclude the case $M = \sum_{k=1}^{K} M_k$ from our analyses since it occurs wirelessly in most GWAS summary datasets.

Denser representation of GPN.

We first define a signed bipartite GPN, $\mathcal{G}_{GPN} = (Y, G, E)$, where $Y = \{Y_1, \cdots, Y_K\}$ and $G = \{G_1, \cdots, G_M\}$ denote two disjoint and independent sets of phenotypes and genetic variants, and E denotes the set of edges in GPN. Similar to Chapter 1, denote $\mathbf{T} = (T_{mk})$ as an $M \times K$ adjacency matrix of GPN, where $T_{mk} = \text{sign}(\hat{\beta}_{mk}) F_{Chi}^{-1}(1 - p_{mk})$ is the weight of the edge between the m^{th} genetic variant and the k^{th} phenotype. $F_{Chi}(\bullet)$ denotes the cumulative distribution function (CDF) of χ_1^2; $\text{sign}(\hat{\beta}_{mk}) = 1$ if $\hat{\beta}_{mk} > 0$, $\text{sign}(\hat{\beta}_{mk}) = 0$ if $\hat{\beta}_{mk} = 0$, otherwise, $\text{sign}(\hat{\beta}_{mk}) = -1$. Note that $|T_{mk}|$ represents the strength of the association and $\text{sign}(\hat{\beta}_{mk})$ represent the direction of the association.

The construction of $\mathbf{T}$ can be considered to be a denser representation of GPN, where $T_{mk} \neq 0$ for $m = 1, \cdots, M_k$ and $k = 1, \cdots, K$. The denser representation includes all associations and does not involve thresholding. Same as the expression quantitative trait locus (eQTL) network construction introduced by Gaynor et al.[31], the denser representation of GPN allows us to capture the fact that we have no prior knowledge of precisely which genetic variants and phenotypes might have an association.

Sparse representations of GPN.

Sparsity makes biological sense as even disease-associated genetic variants are known to generally have a small effect size, meaning they are unlikely to exert their influence across the genome[31]. Therefore, we introduce the false discovery rate (FDR) based sparse representations of GPN, where the networks only include edges where associations meet a measure of significance below a threshold from the denser representation of GPN. Let $\mathcal{G}_{GPN}^{\tau} = (Y, G, E^{\tau})$ be a sparse representation of a bipartite GPN for a specific threshold τ, where E^{τ} denotes the set of edges in the sparse representation of GPN. $\mathbf{T}^{\tau} = (T_{mk}^{\tau})$ is an $M \times K$ adjacency matrix of GPN, where $T_{mk}^{\tau} = T_{mk} \cdot \text{I}(p_{mk}^{*} < \tau)$ with T_{mk} from the denser representation of GPN. $\text{I}(\bullet)$ is an indicator function that takes value 1 when $p_{mk}^{*} < \tau$, otherwise, it takes value 0. p_{mk}^{*} is a measure of the significance of genetic association

between the m^{th} genetic variant and the k^{th} phenotype by correcting for the multiple comparisons in each GWAS summary data. We use q-value[90,91] to define p^*_{mk} in our main analyses, but other definitions can also be used, same as Gaynor et al.[31], such as local FDR (LFDR)[92,93] and an adaptation of Benjamini-Hochberg (BH) FDR[94]. We use different thresholds $\tau \in [0,1]$, where $\tau = 1$ represents the denser representation of GPN since all edges are included; $\tau = 0$ represents the empty network with no edges between genetic variants and phenotypes.

Well-defined sparse representation of GPN.

The choice of the threshold, τ, is very important for the GPN construction. The threshold is a sort of information filter, as decreasing τ, the resulting network will change from a denser network to a very sparse one. An overly denser network can sometimes present challenges in understanding the most biologically informative interactions between genetic variants and phenotypes due to the abundance of information. Conversely, an excessively sparse network may result in the loss of important information. The construction of a well-defined sparse representation of GPN can be presented to determine the optimal threshold ($\hat{\tau}$) of $\mathscr{G}^{\tau}_{GPN}$, which can retain the key information about the interactions between genetic variants and phenotypes[95]. Therefore, we propose an approach to determine the optimal threshold by comparing the network properties with a corresponding random network, including connectivity, centrality, and community structure.

More specifically, we first calculate the network "connectance" for each $\mathscr{G}^{\tau}_{GPN}$, which is defined as the ratio of the number of edges in GPN to the total number of possible edges[96,97]. Mathematically, it can be expressed as: $connectance^{\tau} = \#\{E^{\tau}\} / (M \times K)$, where $\#\{\bullet\}$ is the counting measure, that is, $\#\{E^{\tau}\}$ represents the number of edges included in $\mathscr{G}^{\tau}_{GPN}$. The "connectance" of GPN can provide insight into the structure and functioning between genetic variants and phenotypes. As decreasing τ, the resulting network will change from a denser network ($connectance^{\tau=1} \approx 1$) to a very sparse one ($connectance^{\tau=0} = 0$). For a specific τ, we then construct a corresponding random network by shuffling the edges of the original network $\mathscr{G}^{\tau}_{GPN}$. Let $\mathscr{G}^{random}_{GPN} = (Y, G, E^{random})$ be the corresponding random network, where $conectance^{\tau}$ equals to $conectance^{random}$. We also build an adjacency matrix $\mathbf{T}^{random}$ by keeping the same weights of the edges in E^{τ}. Then, we compute the following network properties of $\mathscr{G}^{\tau}_{GPN}$ and $\mathscr{G}^{random}_{GPN}$, respectively.

Clustering coefficient. The clustering coefficient measures the extent to which two genetic variants share the same set of phenotypes. A genetic variant with a high cluster coefficient indicates that it tends to be associated with similar phenotypes; a genetic variant with a low cluster coefficient indicates that it tends to be associated with distinct sets of phenotypes. Let $N(m)$ and $N(\tilde{m})$ be a set of phenotypes that are linked to the m^{th} genetic variant and the $\tilde{m}^{th}$ genetic variant, respectively. Similar to Latapy et al.[98], we first define the clustering coefficient for any pair of genetic variants $(m, \tilde{m})$ as

$cc(m,\tilde{m}) = \#\{N(m) \cap N(\tilde{m})\} / \#\{N(m) \cup N(\tilde{m})\}$. Let $N^2(m)$ be the sets of genetic variants that are linked to $N(m)$. The genetic variants in $N^2(m)$ are defined as the distance 2 neighbors of the m^{th} genetic variant, that is, two genetic variants are linked together if they have at least one associated phenotype in common. We then define the clustering coefficient for the m^{th} genetic variant as $cc(m) = \sum_{\tilde{m} \in N^2(m)} cc(m,\tilde{m}) / \#\{N^2(m)\}$. Finally, the clustering coefficient of all genetic variants is the average of cluster coefficients of each genetic variant, $cc = \sum_{m=1}^{M} cc(m)/M$. The clustering coefficient of all phenotypes can also be defined in the same way as that of all genetic variants. We also calculate the clustering coefficient of all genetic variants and phenotypes for the random network $\mathscr{G}_{GPN}^{random}$.

Weighted and unweighted degree. The unweighted degree of a genetic variant (phenotype) in a bipartite GPN is defined as the number of edges across all phenotypes (genetic variants)[76]. The unweighted degree of the m^{th} genetic variant and the k^{th} phenotype are defined as $d_m^{G,unweight} = \sum_{k=1}^{K} \mathrm{I}(T_{mk}^{\tau} \neq 0)$ and $d_k^{P,unweight} = \sum_{m=1}^{M} \mathrm{I}(T_{mk}^{\tau} \neq 0)$. The weighted degree is reflecting the strength of the associations of edges, which are defined as $d_m^{G,weight} = \sum_{k=1}^{K} |T_{mk}|$ and $d_k^{P,weight} = \sum_{m=1}^{M} |T_{mk}|$.

Kullback–Leibler (KL) divergence. We define KL divergences[99,100] of degree of genetic variant and phenotypes between $\mathscr{G}_{GPN}^{\tau}$ and $\mathscr{G}_{GPN}^{random}$ to determine the diversities between a bipartite GPN and a random bipartite network, which are given by

$$KL\left(D_{\tau}^{G} \,||\, D_{\tau}^{G,random}\right) = \sum_{m=1}^{M} \bar{d}_m^{G} \log\left(\bar{d}_m^{G} / \bar{d}_m^{G,random}\right),$$

$$KL\left(D_{\tau}^{P} \,||\, D_{\tau}^{P,random}\right) = \sum_{k=1}^{K} \bar{d}_k^{P} \log\left(\bar{d}_k^{P} / \bar{d}_k^{P,random}\right),$$

where $\bar{d}_m^{G}$ and $\bar{d}_k^{P}$ are the min-max standardized degree (either weighted and unweighted) which is defined as $\bar{d}_m^{G} = \left(d_m^{G} - \min_m\{d_m^{G}\}\right) / \left(\max_m\{d_m^{G}\} - \min_m\{d_m^{G}\}\right)$ for the m^{th} genetic variant and $\bar{d}_k^{P} = \left(d_k^{P} - \min_k\{d_k^{P}\}\right) / \left(\max_k\{d_k^{P}\} - \min_m\{d_k^{P}\}\right)$ for the k^{th} phenotype. $KL\left(D_{\tau}^{G} \,||\, D_{\tau}^{G,random}\right)$ and $KL\left(D_{\tau}^{P} \,||\, D_{\tau}^{P,random}\right)$ are used to measure the difference between degree distributions of genetic variants and phenotypes in $\mathscr{G}_{GPN}^{\tau}$ and $\mathscr{G}_{GPN}^{random}$. $KL\left(D_{\tau}^{G} \,||\, D_{\tau}^{G,random}\right)$ will equal 0 if the degree of genetic variants are the same in $\mathscr{G}_{GPN}^{\tau}$ and $\mathscr{G}_{GPN}^{random}$; it will be negative if most degrees in $\mathscr{G}_{GPN}^{random}$ are greater than those in $\mathscr{G}_{GPN}^{\tau}$; and it will be positive if most degrees in $\mathscr{G}_{GPN}^{\tau}$ are greater than those in $\mathscr{G}_{GPN}^{random}$. $KL\left(D_{\tau}^{P} \,||\, D_{\tau}^{P,random}\right)$ has the similar properties. We also define a global KL divergency of a bipartite network as $KL\left(D_{\tau} \,||\, D_{\tau}^{random}\right) = KL\left(D_{\tau}^{G} \,||\, D_{\tau}^{G,random}\right) + KL\left(D_{\tau}^{P} \,||\, D_{\tau}^{P,random}\right)$.

Without loss of the generality, the optimal threshold τ should be selected by maximizing $KL\left(D_\tau^G \,\|\, D_\tau^{G,random}\right)$ and $KL\left(D_\tau^P \,\|\, D_\tau^{P,random}\right)$. Meanwhile, in the case of equivalent numbers and weights of edges in the original network and the corresponding random network, the greater the difference of network topologies between $\mathscr{G}_{GPN}^{\tau}$ and $\mathscr{G}_{GPN}^{random}$, the more information $\mathscr{G}_{GPN}^{\tau}$ includes. Therefore, we also assess the difference for cluster coefficients between $\mathscr{G}_{GPN}^{\tau}$ and $\mathscr{G}_{GPN}^{random}$ for $\tau \in [0,1]$, which is defined as $\Delta cc = cc^{\tau} - cc^{random}$. To investigate the stability of the diversities, $KL\left(D_\tau^G \,\|\, D_\tau^{G,random}\right)$ and $KL\left(D_\tau^P \,\|\, D_\tau^{P,random}\right)$, we construct 1,000 random networks corresponding to $\mathscr{G}_{GPN}^{\tau}$. We thus can estimate the standard error of KL divergence and then obtain the stability by computing their 95% confidence intervals (CIs). For a clustering coefficient, we only evaluate it by generating one random network since it is time consuming in a large-scale network. We also evaluate two other network properties, degree entropy and cross entropy of degree (details in **Text B.1**).

2.2.2 Network topology annotations

For both denser and sparse representations of GPN, we constructed two probabilistic annotations based on the following network centralities. The centralities of a bipartite network are measuring the importance of genetic variants (phenotypes) across phenotypes (genetic variants) in the network. To simplify the notation, we use **T** to denote the adjacency matrix of GPN, which can be constructed by either denser or the sparse representation introduced in section 2.2.1.

***Degree centrality*.**

For the bipartite GPN, a genetic variant with a high degree across phenotypes is more likely to be pleiotropic, as it is highly connected to multiple phenotypes; a phenotype with a high degree across genetic variants is more likely to have higher heritability, as it is connected to multiple genetic variants or with higher association evidence. Same as section 2.2.1, the weighted degree of the m^{th} genetic variant and the k^{th} phenotype are defined as follows:

$$d_m^G = \sum_{k=1}^{K} |T_{mk}| \text{ and } d_k^P = \sum_{m=1}^{M} |T_{mk}|.$$

***Approximate betweenness centrality*.**

In a bipartite GPN, we define an approximate betweenness centrality of a genetic variant which can be used to measure its importance in connecting different phenotypes. A genetic variant with high approximate betweenness can be considered an important connector between phenotypes. The approximate betweenness centrality of the m^{th} genetic variant is defined as

$$b_m = \sum_{(k,l)\in Y} \sigma_{k,l}(m) \Big/ \max\{\sigma_{k,l}, 1\},$$

where $\sigma_{k,l}$ is the number of shortest paths between the k^{th} phenotype and the l^{th} phenotype and $\sigma_{k,l}(m)$ is the number of the shortest path between the k^{th} phenotype and the l^{th} phenotype that pass through the m^{th} genetic variant. Note that there are no direct edges between phenotypes in the bipartite GPN. Therefore, the shortest path $\sigma_{k,l}$ is the number of genetic variants that are associated with both the k^{th} phenotype and the l^{th} phenotype; the shortest path $\sigma_{k,l}(m)$ only takes the value 0 or 1, where $\sigma_{k,l}(m)=1$ if the m^{th} genetic variant is associated with both the k^{th} phenotype and the l^{th} phenotype, otherwise, $\sigma_{k,l}(m)=0$.

2.2.3 Heritability enrichment of network annotations

Note that the network topology annotations of genetic variants quantify the possibility of a genetic variant being a pleiotropic variant. To study whether these annotations are enriched for disease heritability of the highly correlated phenotype, we first perform a leave-one-phenotype-out (LOPO) approach to construct the network topology annotations. Then, we use stratified LD score regression (S-LDSC) to estimate the enrichment and the standardized effect size of an annotation[84,85].

Leave-one-phenotype-out (LOPO).

In this section, we consider K highly genetically correlated phenotypes. To simplify the notation, we use $\tilde{\mathbf{T}}_k$ to denote the adjacency matrix of GPN by removing the k^{th} phenotype in the construction. $\tilde{\mathbf{T}}_k$ can be constructed by either denser or one of the sparse representations introduced in section 2.2.1. Then, we use one of the network topology annotations in section 2.2.2 to assign the numeric value to each genetic variant for the evaluation of the k^{th} phenotype. Assigning a network topology annotation to each genetic variant is a way to quantify its potential for pleiotropy. The LOPO approach can assist in determining whether genetic variants have highly evidenced impacts on other $K-1$ phenotypes through pleiotropy and can also contribute to the heritability of the k^{th} phenotype.

Stratified LD score regression (S-LDSC).

S-LDSC is a method to assess the contribution of the annotation to disease heritability[84,85] conditional on other functional annotations. We use 86 functional annotations in the baseline-LD model (v2.1)[101], including regulatory annotations (e.g., promoter, enhancer, histone marks, TF binding sites), LD-related annotations, et al. In this section, we ignore the index of k to simplify the notations. Let a_{mc} be the annotation value of the m^{th} genetic variant for the c^{th} annotation, where $m=1,\cdots,M_k$ and $c=0,\cdots,C$. In particular, a_{m0} represent the network topology annotation of the m^{th} genetic variant constructed by the LOPO approach.

S-LDSC assumes that the per-SNP heritability or variance of the effect size of each genetic variant is given by $Var(\beta_m) = \sum_{c=0}^{C} a_{mc}\tau_c$, where τ_c is the per-SNP contribution of the c^{th} annotation to disease heritability. We can estimate τ_c using S-LDSC,

$$E(\chi_m^2) = N\sum_{c=0}^{C} l(m,c)\tau_c + 1,$$

where χ_m^2 is the chi-square test statistic for testing the association between the m^{th} genetic variant and a phenotype in GWAS summary data, $l(m,c) = \sum_{\tilde{m}} a_{mc} r_{m,\tilde{m}}^2$ is the LD score of the m^{th} genetic variant to the c^{th} annotation, and $r_{m,\tilde{m}}$ is the genotypic correlation between the m^{th} and the $\tilde{m}^{th}$ genetic variants.

We only focus on the network topology annotation τ_0. As demonstrated by Finucane et al.[102], τ_0 will be positive if the network annotation increases per-SNP heritability, accounting for all other factors. Let $sd(\boldsymbol{a}_0)$ be the standard deviation of the network topology annotation. The standardized effect size τ_0^* is defined by

$$\tau_0^* = \frac{\tau_0 sd(\boldsymbol{a}_0)}{\sum_m Var(\beta_m)/M_k}.$$

Note that τ_0^* is defined as the proportionate change in per-SNP heritability associated with a one-standard-deviation increase in the network topology annotation conditioning on all the other annotations[85]. The standard error on the estimate of τ_0^*, $sd(\tau_0^*)$, is computed using a block jackknife[84]. Then, we can compute the p-value to test if $\tau_0^* > 0$ by assuming $\tau_0^*/sd(\tau_0^*) \sim N(0,1)$.

We also calculate the enrichment of the network topology annotation, which is defined as the proportion of the heritability explained by genetic variants in the annotation divided by the proportion of genetic variants in the annotation.

$$Enrichment = \frac{h_g^2(\tau_0)/h_g^2}{\sum_m a_{m0}/M_k},$$

where $h_g^2 = \sum_m Var(\beta_m)$ is the estimated heritability and $h_g^2(\tau_0)$ is the heritability captured by the network annotation. $Enrichment > 1$ represents the network annotation is not enriched for the disease heritability. Same as τ_0^*, the significance for *Enrichment* is computed using a block jackknife[84]. The inclusion of the 86 functional annotations in the baseline-LD model can minimize the risk of bias in enrichment estimates and can also estimate the effect size τ_0 conditional on known functional annotations[84].

2.2.4 Community detection methods

Community detection methods are essential in comprehending the global and local structures of associations between genetic variants and phenotypes, and in shedding light on association connections that may not be easily visible in the network topology[33]. Calculating the projection of GPN onto phenotypes that are linked through shared genetic variants is a very important step in the community detection. Let $\mathcal{G}_{PPN} = \left(Y, E^P\right)$ be the one-mode projection of GPN, called Phenotype and Phenotype Network (PPN), where E^P denotes the set of edges between phenotypes in PPN. Denote $\mathbf{W} = \left(W_{kl}\right)$ as an $K \times K$ adjacency matrix of PPN, where W_{kl} is the weight of the edge between the k^{th} phenotype and the l^{th} phenotype. In this study, we perform the community detection methods to partition K phenotypes into L disjoint network modules based on the adjacency matrix of PPN.

Community detection method for the denser representation of GPN.

For the denser representation of GPN, one straightforward way to define the adjacency matrix $\mathbf{W}$ is to compute the direct correlation of $\mathbf{T}$, $\mathbf{W} = cor\left(\mathbf{T}\right)$[77]. The elements of $\mathbf{W}$ can be both positive and negative, indicating that the PPN given by the adjacency matrix of $\mathbf{W}$ is a signed network. Inspired by our previously proposed modularity-based community detection method[103], we introduce a community detection method for the signed network in this study. Let $\mathbf{W}^+ = \left(W_{kl}^+\right)$ and $\mathbf{W}^- = \left(W_{kl}^-\right)$ be adjacency matrices of the positive and negative weights, respectively, where $W_{kl}^+ = \max\left\{W_{kl}, 0\right\}$ and $W_{kl}^- = -\min\left\{W_{kl}, 0\right\}$ such that $\mathbf{W} = \mathbf{W}^+ - \mathbf{W}^-$. First, we assume K phenotypes can be divided into k_0 network modules using a hierarchical clustering method with similarity matrix $\mathbf{W}$ for $k_0 = 1, \cdots, K$. Let $\mathbf{C}^{(k_0)} = \left(C_{k,l}^{(k_0)}\right)$ be a $K \times K$ connectivity matrix, where $C_{k,l}^{(k_0)} = 1$ if the k^{th} phenotype and the l^{th} phenotype are in the same network module, otherwise, $C_{k,l}^{(k_0)} = 0$. Then, we calculate the modularity of network with only positive weights, $\mathbf{W}^+$, for each k_0 as $Q_{k_0}^+ = \frac{1}{2D^+}\sum_{k,l=1}^{K}\left(W_{kl}^+ - \frac{d_k^+ d_l^+}{2D^+}\right)C_{k,l}^{(k_0)}$, where $d_k^+ = \sum_{l=1}^{K} W_{kl}^+$ and $D^+ = \sum_{k=1}^{K} d_k^+ \big/ 2$ represent the degree of the k^{th} phenotype and overall degree of $\mathbf{W}^+$. Similarly, we calculate the modularity of $\mathbf{W}^-$ as $Q_{k_0}^-$. Therefore, we define the modularity for the signed network as $Q_{k_0} = \frac{2D^+}{2D^+ + 2D^-}Q_{k_0}^+ - \frac{2D^-}{2D^+ + 2D^-}Q_{k_0}^-$. Note that a network with a high modularity value has dense connections between phenotypes within network modules but sparse connections between phenotypes in different modules[33]. Then, we determine the optimal number of network modules as $L = \arg\max\left\{Q_1, Q_2, \cdots, Q_K\right\}$.

Community detection method for the sparse representation of GPN.

To eliminate the biases in projections caused by a large number of genetic variants that are unlikely to exert their influence across the whole genome[31], we also provide a weighted projection approach by only focusing on the shared genetic variants between two phenotypes in the (well-defined) sparse representations of GPN, $\mathbf{T}^{sparse}$. Let S_{kl}^{*} be the set of the genetic variants that are connected with the k^{th} phenotype and the l^{th} phenotype. We define $W_{kl}=\sum_{m\in S_{kl}^{*}}\left|T_{mk}^{sparse}\right|\Big/ d_{k}^{sparse}$ and $W_{lk}=\sum_{m\in S_{kl}^{*}}\left|T_{ml}^{sparse}\right|\Big/ d_{l}^{sparse}$, where d_{k}^{sparse} and d_{l}^{sparse} are the weighted degree of the k^{th} and the l^{th} phenotypes, respectively. More specifically, W_{kl} is a proportion of degree of the k^{th} phenotype explained by the shared associations between the k^{th} and the l^{th} phenotypes; similarly, W_{lk} is a proportion of degree of the l^{th} phenotype explained by the shared associations between the k^{th} and the l^{th} phenotypes. Therefore, $W_{kl}\neq W_{lk}$ indicates that the projected PPN is a directed network. If $W_{kl}>W_{lk}$, the shared associations between the k^{th} and the l^{th} phenotypes are more important to the k^{th} phenotype than the l^{th} phenotype. In particular, $W_{kl}=1$ if and only if the k^{th} phenotype only links with the genetic variants in S_{kl}^{*}. The modularity is easily generalized to both weighted and directed network, where the modularity based on LinkRank is given by[104,105]: $Q_{k_0}=\sum_{k,l=1}^{K}\left(\pi_k G_{k,l}-\pi_k\pi_l\right)C_{k,l}^{(k_0)}$. Let $W_{k}^{out}=\sum_{l=1}^{K}W_{kl}$ be the out-degree of the k^{th} phenotype for a directed PPN. Then, $\pi_1,\cdots,\pi_K$ is the PageRank vector indicating the probability of a phenotype being visited by a random surfer. $G_{k,l}=\alpha\cdot W_{kl}\big/W_{k}^{out}+1\big/K\cdot\left(\alpha g_k+1-\alpha\right)$ is the Google Matrix, where α is the damping parameter for PageRank[104] (with probability $1-\alpha$ random surfer jumps to a random phenotype) and $g_k=\mathrm{I}\left(W_{k}^{out}=0\right)$ is an indicator of dangling phenotype. Same as the community detection method for the denser representation of GPN, we also determine the optimal number of network modules as $L=\arg\max\left\{Q_1,Q_2,\cdots,Q_K\right\}$.

2.2.5 Phenome-wide association studies (PheWAS)

The community detection method for PPN based on $\mathbf{W}$ has potential applications in PheWAS and multiple phenotype association studies. In Chapter 1, we introduced the application of multiple phenotype association tests for analyzing K correlated phenotypes. In this section, we extend our discussion to include the application of GPN and PPN in PheWAS. We can obtain an a priori grouping of phenotypes from the community detection method of GPN and PPN, then jointly test multiple phenotypes in each network module and one genetic variant to discover the cross-phenotype associations and pleiotropy.

Assume that K is the total number of phenotypes in the whole phenome, which can be partitioned into L disjoint network modules from section 2.2.4. Let $K=K_1+\cdots+K_L$, where K_l is the number of phenotypes in the l^{th} network module. In this section, we apply four powerful GWAS summary-based multiple phenotype association

tests to identify the association between phenotypes in the l^{th} network module and a genetic variant, including minP[17], ACAT[106], MTAG[107], SHom[108] (details in **Text B.2**). Then, we refine our previous approach to evaluate FDR by thresholding the p-values obtained from the multiple phenotype association tests[47]. Let $\left\{ p_m^{(1)}, \cdots, p_m^{(L)} \right\}$ be a sequence of p-values for testing the association between phenotypes in each of the network modules and the m^{th} genetic variant. For a given nominal FDR level $\alpha \in (0,1)$, the optimal threshold for the m^{th} genetic variant is given by

$$\hat{p}_m = \sup\left\{ p \in [0,1] : t \le \frac{\alpha \max\left\{1, \sum_{l=1}^{L} \mathrm{I}\left(p_m^{(l)} \le p \right)\right\}}{m_0} \right\},$$

where m_0 is the number of network modules under the null hypothesis that phenotypes in the network module and the m^{th} genetic variant have no association. We refine the estimation $m_0 = L - m_1$, where $m_1 = \sum_{l=1}^{L} \mathrm{I}\left(p_m^{(l)} \le 0.05/L \right)$ is the number of identified network modules that are associated with the m^{th} genetic variant based on the Bonferroni Correction.

2.2.6 Empirical GWAS summary datasets

In our analyses, we consider two publicly available GWAS summary datasets to evaluate the performance of constructions of bipartite GPN, heritability enrichment of network annotations, community detection methods, and applications of PheWAS.

***GWAS summary statistics for correlated phenotypes*.**

To perform the heritability enrichment of network annotations, we obtain publicly available GWAS summary data for 12 highly genetically correlated phenotypes in individuals of European ancestry, including attention deficit/hyperactivity disorder (ADHD), smoking initiation (SmkInit), autism spectrum disorder (ASD), neuroticism (NSM), anxiety disorder (AXD), major depressive disorder (MDD), obsessive-compulsive disorder (OCD), anorexia nervosa (AN), bipolar disorder (BD), schizophrenia (SCZ), educational attainment (EA), and cognitive performance (CP). The details of GWAS summary data for the 12 phenotypes are summarized in **Table B.1**. As demonstrated by Zhang et al.[109], the global genetic correlations among the 12 phenotypes estimated by their proposed SUPERGNOVA are ranging from -0.41 to 0.69. 51 out of 66 pairs of phenotypes have significant non-zero global genetic correlations (right upper triangle of **Table B.2**). Meanwhile, they also reported the proportions of correlated regions between two phenotypes that are ranging from 0.11% to 93%. 46 pairs of phenotypes contain at least one significantly correlated region after Bonferroni correction (left lower triangle of **Table B.2**). We only include the genetic variants in 22 autosomes.

***GWAS summary in the UK Biobank*.**

The UK Biobank is a population-based cohort study with a wide variety of genetic and phenotypic information[52]. It recently released genome-wide association data on ~ 500K

individuals from all around the United Kingdom[41,53]. We obtain the publicly available GWAS summary data for 633 EHR-derived phenotypes with main ICD10 diagnoses from Neale lab (Data availability). These GWAS summary data are calculated based on a basic association test on ~337,000 unrelated individuals of British ancestry. We run the LD score regression (LDSC)[110] to each of these 633 phenotypes, therefore, we exclude 45 phenotypes in our analyses since the estimators of their heritability are out of bounds. There are 588 phenotypes across 1,096,648 genetic variants in autosomes in our analyses.

2.3 Results

2.3.1 Construction of GPNs for 12 genetically correlated phenotypes

We construct three bipartite GPNs for 12 genetically correlated phenotypes listed in **Table B.1**, including a denser representation, an arbitrary sparse representation, and a well-defined representation. There are a total of 17,585,432 unique genetic variants from 12 GWAS summary datasets. The global genetic correlations and proportions of correlated regions among the 12 phenotypes estimated by SUPERGNOVA[109] are shown in **Table B.2**. We also perform LDSC[110] to estimate phenotypic correlation (right upper triangle of **Table B.3**) and genetic correlation (left lower triangle of **Table B.3**) among the 12 phenotypes. Among a total of 66 pairs of phenotypes, 45 pairs of phenotypes have significant non-zero genetic correlations ($\text{p-values} < 0.05/66 = 7.58\times10^{-4}$). In particular, MDD has significant non-zero genetic correlations with other 11 phenotypes; NSM has significant non-zero genetic correlations with 10 phenotypes except for BD; SCZ and EA have significant non-zero genetic correlations with other 10 phenotypes, but SCZ and EA do not have significant non-zero genetic correlation.

The denser representation of GPN is constructed without using any thresholds. Since the 12 GWAS summary datasets contain differet numbers of the 17,585,432 unique genetic variants, the connectance of the denser representation of GPN is 0.5123 (**Figure B.1(a)**). The well-defined sparse representation of GPN is constructed by comparing the network properties with the corresponding random networks. Since we only have 12 phenotypes in this analysis, we only consider the network properties for genetic variants of the constructed GPN and the corresponding random networks. For each $\tau \in (0,1)$, we generate 1,000 corresponding random networks. **Figure 2.2(a)** shows the comparisons of the KL divergence for genetic variants across 1,000 random networks. The KL divergence increases from 0 to a specific value of the threshold and then decreases from that value to 1, indicating that the difference between the original and random network reaches the maximum at the specific value. We also calculate the cross entropy of the weighted degree of genetic variants compared to the corresponding random network (**Figure 2.2(b)**).

Note that the weighted degree of genetic variants in a corresponding random network becomes more different than the original one if the original network retains the key information about the interactions between genetic variants[95]. The network properties, KL divergence and cross entropy, will reach the maximum value at the most informative network. In our analysis, we prioritize choosing the optimal threshold with respect to KL divergence and then check the cross entropy and weighted degree entropy at that optimal threshold. The maximum mean of KL divergence equals 9.02×10^{8} at $\tau = 0.45$, where the

mean of cross entropy equals a larger value (9.83×10^5) even though it does not reach the maximum value. Therefore, we constructed the well-defined sparse representation of GPN with $\tau = 0.45$. This optimal threshold is much larger than the significant level for the association testing (e.g., $\tau = 0.05$ for controlling FDR at the nominal level of 0.05). The optimal threshold in the construction of GPN does not represent the significant associations between genetic variants and phenotypes. It is only used to ensure that the constructed GPN is more informative than a random network.

As a comparison, we also construct an arbitrary sparse representation of GPN by using the threshold $\tau = 0.1$. **Figure 2.2(c)** shows the weighted degree distribution of genetic variants for three GPNs, denser representation ($\tau = 1$), well-defined sparse representation ($\tau = 0.45$), and an arbitrary threshold sparse representation ($\tau = 0.1$). We observe that the degree distributions of all three networks follow the power law with different scale parameters γ, indicating that a small number of genetic variants have a much larger number of connections than the majority of genetic variants. In particular, the degree of genetic variants in the denser representation of GPN is greater than those in a sparser GPN, resulting in the scale parameter increases with increasing the threshold τ.

We also calculate the network properties of the unweighted GPNs by comparing them with the corresponding random networks (**Figure B.2**). Furthermore, the adjacency matrix of the projected PPN, **W** introduced in section 2.2.4, can be considered as the phenotypic correlation among 12 phenotypes based on the shared genetic architecture. **Figure B.3** shows the comparisons of the adjacency matrix of PPN constructed by the denser and well-defined sparse representations of GPN with the genetic correlation matrix estimated by SUPERGNOVA[109] (**Table B.2**) and LDSC[110] (**Table B.3**).

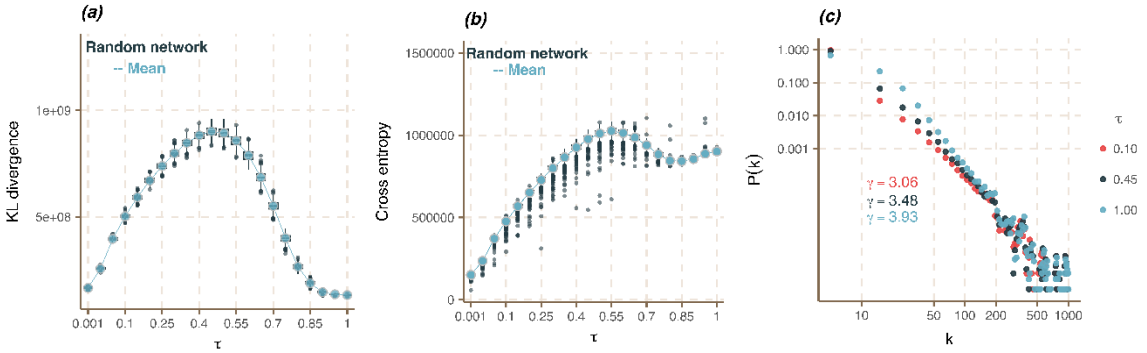


Figure 2.2. Network properties of the weighted bipartite GPNs for 12 genetically correlated phenotypes. (a) KL divergency for genetic variants. The blue line is the mean of KL divergencies across 1,000 random network comparisons. The boxplots show the scaled distributions of the KL divergency for each threshold. (b) Cross entropy for genetic variants. Blue lines are the means of cross entropy across 1,000 random network comparisons. The boxplot shows the scaled distribution of the cross entropy for each threshold. (c) Plot of the weighted degree distribution of genetic variants for three GPNs on the log-log scale, denser representation ($\tau = 1$), well-defined sparse representation ($\tau = 0.45$), and an arbitrary threshold sparse representation ($\tau = 0.1$).

2.3.2 Heritability enrichment analysis of network annotations

For each of the above three bipartite GPNs for 12 phenotypes, we perform S-LDSC along with LOPO to evaluate whether the network topology annotations are enriched for disease heritability. We consider both degree centrality and betweenness centrality of genetic variants, conditioning on 86 functional annotations in the baseline-LD model (v2.1)[101]. These 86 existing functional annotations have been demonstrated to be highly informative by capturing functionality and LD-related features, thus, we evaluate the added value of our network topology annotations in capturing disease heritability, contributed by the pleiotropic variants with other genetically correlated phenotypes.

Table 2.1 shows the heritability enrichment analysis results for degree centrality calculated from denser, arbitrary sparse, and well-defined sparse representations of GPN, respectively. From the LDSC results (**Table B.3**), MDD has significant non-zero genetic correlations with all other 11 phenotypes. **Table 2.1** shows that the degree centrality annotation is significantly enriched for the heritability of phenotype MDD based on all of the three constructed GPNs ($\text{p-values} < 0.05/12 \approx 0.0042$). As we demonstrated in section 2.2.2, the network topology annotation of each genetic variant quantifies its possibility for pleiotropy among other correlated phenotypes. After we use the LOPO approach to construct the network annotations of MDD, the significance enrichment indicates that the network annotation can contribute more information to disease heritability if it is computed based on other highly genetically correlated phenotypes. In particular, even though the arbitrary sparse representation of GPN ($\tau = 0.1$) contains less information than the denser and well-defined GPN, the degree centrality annotation is still significantly enriched in heritability of MDD ($\text{p-value} = 2.79 \times 10^{-5}$) conditioned on the 86 functional annotations. Meanwhile, the degree annotation is also significantly enriched in heritability of CP ($\text{p-value} = 2.76 \times 10^{-6}$) and SCZ ($\text{p-value} = 0.0021$) for the arbitrary sparse representation of GPN: SCZ has significant non-zero genetic correlations with 10 phenotypes except for EA (**Table B.3**); CP has the significant proportions of correlated regions with 9 phenotypes in which there are over 15% of correlated regions with 8 phenotypes (**Table B.2**).

The network annotation based on degree centrality obtained by the denser representation of a bipartite GPN includes the complete information for explaining the associations between phenotypes and genetic variants. It is significantly enriched to disease heritability of 11 out of 12 phenotypes as expected, except for AXD, with enrichment estimates ranging from 1.4457 (OCD with $\text{p-value} = 0.0016$) to 2.2894 (ASD with $\text{p-value} = 8.69 \times 10^{24}$). We identify the most significant enrichment of network annotations based on degree centrality for CP ($\text{Enrichment} = 2.2026$ with $\text{p-value} = 6.33 \times 10^{-54}$) and EA ($\text{Enrichment} = 2.0406$ with $\text{p-value} = 1.14 \times 10^{-52}$). These two phenotypes have a significant proportion of correlated regions, 93%, estimated by SUPERGNOVA[109]. **Figures B.4(a) and B.4(b)** show the qq-plot of EA versus CP based on the weight of the denser and the well-defined sparse representations of GPN. Most of the genetic variants have similar weights for both EA and CP, lying in the diagonal line, but there exist some genetic variants that have the largest weights for only one phenotype. The same

relationship between EA and CP is shown in the marginal associations from GWAS summary datasets (**Figures B.4(c) and Figure B.4(d)**).

The network topology annotations obtained by the well-defined sparse representation of GPN ($\tau = 0.45$) perform similarly on the heritability enrichment compared to the denser representation of GPN. Even though some information is excluded from the well-defined GPN, the annotations obtained by the well-defined GPN contribute similar effects to disease heritability. **Table 2.1** and **Table B.4** show the annotations from both denser and well-defined sparse representations of GPN can significantly enrich to disease heritability of the same phenotypes. However, the network topology annotations obtained by the arbitrary sparse representation of GPN ($\tau = 0.1$) are not enriched to most disease heritability. We can conclude that a more informative network can be used to understand heritability rather than an arbitrary one with a smaller threshold. For example, if we use the significance level of the associations (e.g., $\tau = 0.1$ or $\tau = 0.05$) to construct a GPN, it may loss more information and key connections even though its edges represent the significant associations between genetic variants and phenotypes.

However, the network annotation based on approximate betweenness centrality performs differently on the heritability enrichment analysis with the annotation based on degree centrality. **Table B.4** shows the heritability enrichment analysis results for betweenness centrality calculated from denser, arbitrary sparse, and well-defined sparse representations of GPN, respectively. We observe that the betweenness centrality calculated by the denser representation of GPN significantly enriches the disease heritability of only seven phenotypes, whereas the annotation calculated by the well-defined GPN can significantly enrich the heritability of 10 phenotypes. The strength of the associations between genetic variants and phenotypes is not considered in the betweenness centrality and the denser representation of GPN includes all edges. Therefore, the betweenness centrality of GPN is not an important feature that can be considered in the heritability enrichment analysis. Alternatively, it is an important network property for the sparse representation of GPN since only the edges with strength evidence of associations are included in the GPN. A genetic variant with high approximate betweenness can be considered an important connector between phenotypes. Therefore, the network annotations based on the approximate betweenness centrality calculated from the well-defined ($\tau = 0.45$) and the arbitrary ($\tau = 0.1$) sparse representation of GPN are significantly enriched to 10 phenotypes' heritability. Meanwhile, the network annotation calculated by a well-defined GPN has stronger evidence than that calculated by the arbitrary one.

According to heritability enrichment results, we observe that network annotations are not enriched to the disease heritability of AXD and OCD. **Figure B.5** shows the heatmap of edge weights in the well-defined sparse representation of GPN for the top 100 and the top 1000 genetic variants with the highest degree centrality, respectively. We observe that these top genetic variants have smaller weights on AXD and OCD, which means that the genetic variants with the highest degree centrality are not associated with AXD and OCD. Therefore, the network annotation is not enriched to their heritability. In particular, there are no edges between OCD and genetic variants if the threshold is smaller than 0.4.

Table 2.1. Heritability enrichment analyses of network topology annotation (degree centrality) based on denser and sparse representations of bipartite GPN for each of the 12 phenotypes.

Trait	Denser		Sparse ($\tau = 0.45$)		Sparse ($\tau = 0.1$)	
	Enrichment (Standard error) *p-value*	Effect τ^* ($se(\tau^*)$) *z-score*	Enrichment (Standard error) *p-value*	Effect τ^* ($se(\tau^*)$) *z-score*	Enrichment (Standard error) *p-value*	Effect τ^* ($se(\tau^*)$) *z-score*
ADHD	2.2175 (0.1697) ***8.26e-24***	3.5434 (0.3247) *10.8870*	3.3012 (0.3209) ***8.49e-22***	3.5192 (0.3423) *10.2797*	3.4734 (0.9173) *0.0072*	2.6504 (0.9882) *2.6820*
AN	1.7796 (0.1097) ***4.31e-21***	1.5274 (0.1694) *9.0145*	2.5216 (0.2174) ***3.73e-19***	1.5866 (0.1823) *8.7030*	2.5594 (0.9810) *0.1119*	1.1405 (0.7423) *1.5364*
ASD	2.2894 (0.2640) ***8.69e-24***	2.2771 (0.2373) *9.5973*	3.4316 (0.4836) ***6.52e-21***	2.3124 (0.2580) *8.9614*	6.1025 (1.9961) *0.0118*	3.5573 (1.4359) *2.4773*
AXD	1.5678 (0.5801) *0.0754*	0.2486 (0.1613) *1.5382*	2.1892 (1.1815) *0.0653*	0.2913 (0.1703) *1.7102*	5.6798 (5.0946) *0.2467*	0.7908 (0.6693) *1.1816*
BD	2.0745 (0.1184) ***7.61e-31***	3.8595 (0.3194) *12.0837*	3.2647 (0.2417) ***1.25e-30***	4.3352 (0.3547) *12.2213*	2.9583 (0.7146) *0.0043*	2.5911 (0.9309) *2.7835*
CP	2.2026 (0.0562) ***6.33e-54***	3.4031 (0.1680) *20.2517*	3.9373 (0.1260) ***2.63e-55***	4.1757 (0.1972) *21.0983*	4.6075 (0.7325) ***2.76e-06***	3.3237 (0.6999) *4.7485*
EA	2.0406 (0.0459) ***1.14e-52***	1.9705 (0.1001) *19.5241*	3.7963 (0.1204) ***1.24e-50***	2.4471 (0.1267) *19.3187*	3.5526 (0.8799) *0.0045*	1.2735 (0.4486) *2.8389*
MDD	1.9550 (0.0715) ***4.40e-32***	0.7342 (0.0580) *12.6561*	3.0106 (0.1537) ***1.19e-29***	0.7761 (0.0615) *12.1223*	3.6246 (0.6172) ***2.79e-05***	0.6783 (0.1609) *4.2153*
NSM	1.8706 (0.1088) ***1.06e-19***	1.0423 (0.1147) *9.0888*	2.8629 (0.2225) ***9.01e-20***	1.1485 (0.1243) *9.2426*	4.1886 (1.0518) *0.0097*	1.3055 (0.5086) *2.5669*
OCD	1.4457 (0.2218) ***0.0016***	1.3711 (0.5976) *2.2942*	1.8569 (0.4276) ***0.0022***	1.4454 (0.6231) *2.3197*	0.6951 (2.1090) *0.8867*	-0.5192 (3.1212) *-0.1663*
SCZ	1.9353 (0.0668) ***2.65e-36***	5.4211 (0.3765) *14.3994*	3.0742 (0.1512) ***1.38e-33***	5.6948 (0.4217) *13.5116*	3.2212 (0.7209) ***0.0021***	4.0283 (1.3343) *3.0190*
SmkInit	1.6750 (0.0918) ***9.76e-21***	0.5857 (0.0675) *8.6809*	2.3947 (0.1866) ***8.62e-20***	0.6398 (0.0731) *8.5610*	2.1556 (0.8704) *0.1839*	0.3691 (0.2839) *1.2888*

Notes: The estimated effect size and its estimated standard error, τ^* and $se(\tau^*)$, are scaled by dividing 10^{-9}. Z-score of the effect size is reported to test the null hypothesis that either $\tau \leq 0$ (one-sided) or $\tau = 0$ (two-sided). P-value of enrichment is reported to test the null hypothesis that $Enrichment > 1$. The bold-faced p-values indicate the annotation significantly enriched in the disease heritability after accounting for multiple testing (p-value$< 0.05/12 \approx 0.0041$).

2.3.3 Construction of GPNs for 588 EHR-derived phenotypes in the UK Biobank

For a total of 1,096,648 genetic variants and 588 EHR-derived phenotypes with main ICD10 diagnoses after preprocessing, we construct two bipartite GPNs including a denser representation and the well-defined sparse representation. Different from the previous 12 GWAS summary datasets obtained from different studies, GWAS summary datesets of these 588 phenotypes are calculated based on a basic association test on the same ~337,000 unrelated individuals of British ancestry. Therefore, connectance of the denser representation of GPN equals 1, that is, all genetic variants link with all phenotypes with strength of the associations (**Figure B.1(b)**).

We consider the network properties for both genetic variants and phenotypes of constructed GPN and the corresponding random networks. For each $\tau \in (0,1)$, we generate 1,000 corresponding random networks. **Figure 2.3(a) and 2.3(b)** show the KL divergence for genetic variants and phenotypes across 1,000 random network comparisons, respectively. The KL divergence increases from 0 to a specific value of the threshold and then decreases from that value to 1, indicating that the difference between the original and random network reaches the maximum at the specific value. We also calculate the cross entropy and degree entropy of the weighted degree of genetic variants compared to the corresponding random network (**Figure B.6**). The maximum mean of KL divergence equals 1.14×10^{8} at $\tau = 0.6$, where the mean of cross entropy equals 3.90×10^{4} with the largest standard error (17.08) compared with other thresholds. Therefore, we constructed the well-defined sparse representation of GPN with $\tau = 0.6$. We also compare degree distributions of the well-defined network with a more denser representation ($\tau = 0.8$) and two arbitrary threshold sparse representations ($\tau = 0.2$ and $\tau = 0.4$) of GPN. Similar to the constructed GPN of 12 genetically correlated phenotypes, the degree distributions of all four networks follow the power law with different scale parameters γ, indicating that a small number of genetic variants have a much larger number of connections than the majority of genetic variants. In particular, the degree of genetic variants in the denser representation of GPN is greater than those in a sparser GPN, resulting in the scale parameter increases with increasing the threshold τ. Meanwhile, we calculate the network properties of the unweighted GPNs by comparing them with the corresponding random networks (**Figure B.7**).

We calculate three network topology annotations of genetic variants in the constructed GPNs with $\tau = 0.2, 0.4, 0.6, 0.8$, including weighted degree centrality, unweighted degree centrality, and approximate betweenness centrality (**Figures B.7 and B.8**). **Figure B.7** illustrates the relationship between the approximate betweenness centrality of genetic variants and the weighted degree centrality of genetic variants. We mark the genetic variants with the highest centralities.

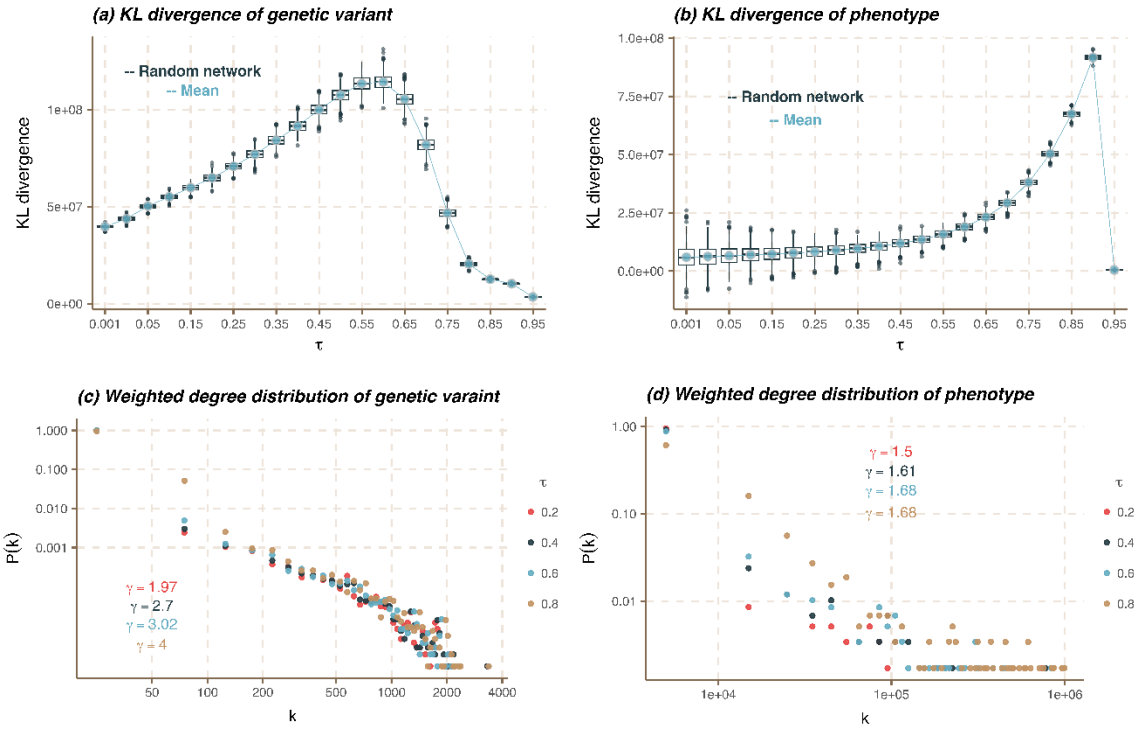


Figure 2.3. Network properties of the bipartite GPNs for 588 EHR-derived phenotypes in the UK Biobanks. (a) and (b) KL divergency for genetic variants and phenotypes. The blue line is the mean of KL divergencies across 1,000 random network comparisons. The boxplots show the scaled distribution of KL divergency for each threshold. (c) and (d) Weighted degree distribution of genetic variants and phenotypes for four GPNs on log-log scale, more denser representation ($\tau = 0.8$), well-defined sparse representation ($\tau = 0.6$), and two arbitrary threshold sparse representations ($\tau = 0.2$ and $\tau = 0.4$).

2.3.4 Community detection for phenotypes

For the denser representation of GPN, we construct the one-mode projected PPN by taking the correlation of the adjacency matrix of GPN. After applying the modularity-based community detection method to the signed PPN, we partition 588 EHR-derived phenotypes into 132 disjoint network modules. The number of phenotypes in each network module ranges from 1 to 87. For the well-defined sparse representation of GPN with, we also construct a directed PPN by only focusing on the shared genetic variants between two phenotypes. In the sparse representation of GPN, phenotypes link with multiple genetic variants, but different phenotypes may not share a link with the same genetic variants. That is, we define the adjacency matrix for the k^{th} phenotype as $W_{kl} = 0$ for all $l = 1, \cdots, K$ if the k^{th} phenotype does not share the same genetic variants with other phenotypes. Therefore, we first isolate 125 phenotypes without sharing any genetic variants with other phenotypes as 125 network modules for a single phenotype. Then, we partition the remaining 463 phenotypes into 71 network modules using the community detection method introduced in section 2.2.4. The number of phenotypes in the 71 network modules

ranges from 2 to 37, and there are a total of 196 network modules. For a comparison, we also apply our proposed community detection method based on denser representation of GPN to LDSC phenotypic correlation. 588 phenotypes are divided into 114 categories with the number of phenotypes ranging from 2 to 82.

2.3.5 Phenome-wide association studies (PheWAS)

In PheWAS, a priori grouping (network module) of phenotypes in whole phenome can be obtained by the community detection of PPN. For each network module, we jointly test the phenotypes within this module and a genetic variant to discover the cross-phenotype associations and potential pleiotropy. In this study, we perform five powerful GWAS summary-based multiple phenotype association tests to identify the association between phenotype in each network module and each of genetic variants, including minP[17], ChiSq[17], ACAT[106], MTAG[107], SHom[108] (details in **Text B.2**). Then, we use the refined FDR controlling approach to evaluate FDR by thresholding the p-values obtained from the multiple phenotype association tests.

Simulation studies.

We first conduct extensive simulation studies to evaluate whether multiple phenotype association tests used in our study can well-control FDR. We consider two simulation settings of the number of phenotypes: 500 phenotypes with 50 phenotypic categories and 1,000 phenotypes with 100 phenotypic categories (details in **Text B.3**). We assume that only the phenotypes within the same phenotypic category are correlated with each other. Similar to Lee et al.[9], we consider two scenarios of correlations among phenotypes within the same category: 1) same correlation between each pair of phenotypes (SAME); 2) different correlation between each pair of phenotypes that is generated by using an autoregressive (AR(1)) model. **Table B.4** and **Table B.5** show the average FDR in the simulation studies for 500 phenotypes and 1,000 phenotypes, respectively. FDR is evaluated using 10 Monte-Carlo (MC) runs, equivalent to 1,000 replications at a nominal FDR level of 5% (**Text B.3**). The 95% confidence interval (CI) is (0.0365, 0.0635). Note that we directly generate z-scores instead of effect sizes of genetic variants on phenotypes without considering LD, therefore, we do not consider MTAG in our simulation studies. The correlations among phenotypes are estimated by the method introduced in Kim et al.[17]. We observe that minP cannot control FDR in all scenarios but ACAT, and SHom well-control FDR as expected.

PheWAS based on 165 UK Biobank level 1 categories:

As benchmarked categories in our analysis, we use 165 UK Biobank level 1 categories defined in data-field 41202 (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41202). The number of phenotypes in each category ranges from 1 to 20: there are 43 categories containing only one phenotype; 35 and 31 categories contain 2 and 3 phenotypes, respectively; only 7 categories contain more than 10 phenotypes. In our real data analyses, we only apply three multiple phenotype association tests (ACAT, SHom, and MTAG) to test the association between phenotypes in each network module and each genetic variant. minP is not considered here since it cannot control FDR evaluated in our simulation studies. We use the commonly used genome-wide nominal FDR level 5×10^{-8}. After applying our refined FDR controlling approach for the tests of each genetic variant, ACAT can identify

6,105 genetic variants associated with at least one category. We observe that most genetic variants are associated with only one category. SHom can identify 2,701 genetic variants and MTAG can identify 2,980 genetic variants (**Figure 2.4**).

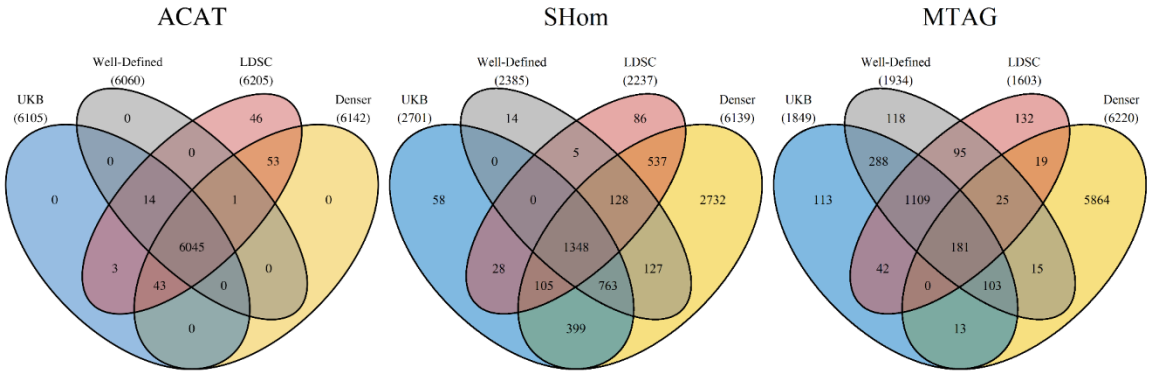


Figure 2.4. Venn plots for genetic variants identified by three multiple phenotype association tests based on different phenotypic categories and network modules.

PheWAS based on 114 phenotypic categories from LDSC.

As a comparison, we also apply three multiple phenotype association tests to 114 categories detected from the phenotypic correlation estimated by LDSC. ACAT identifies 6,205 genetic variants, SHom identifies 2,237 genetic variants, and MTAG identifies 1,603 genetic variants. Compared with the association tests based on the phenotypic categories in the UK Biobank, ACAT based on the LDSC can identify all of the 6,105 genetic variants identified by ACAT based on the UK Biobank (**Figure 2.4**). Meanwhile, there are 100 genetic variants that are uniquely identified by ACAT based on the LDSC. **Figure B.10** shows the heatmap of -log10(p-value) from GWAS summary datasets of these 100 genetic variants. We observe that all of these 100 genetic variants significantly associated with at least one phenotype at the GWAS significance level 5×10^{-8}. According to results from SHom and MTAG, tests based on the UK Biobank identify more genetic variants than the tests based on the LDSC.

PheWAS based on 132 network modules from denser representation of GPN

According to the 132 network modules from denser representation of GPN (section 2.3.4), ACAT can identify 6,142 genetic variants associated with at least one network modules and SHom can identify 6,139 genetic variants. In the application of MTAG, it is time consuming and out of memory for one network module with 87 phenotypes. Therefore, we perform MTAG on the other 131 network modules and MTAG identifies 6,220 genetic variants. **Figure 2.4** shows the Venn plot for genetic variants identified by three multiple phenotype association tests based on different phenotypic categories and network modules. Based on the network modules detected from the denser representation of GPN, all three methods (ACAT, SHom, and MTAG) can identify ~6,000 genetic variants associated with at least one network module.

PheWAS based on 196 network modules from well-defined representation of GPN.

According to the 196 network modules from well-defined representation of GPN (section 2.3.4), ACAT can identify 6,060 genetic variants associated with at least one network modules; SHom can identify 2,385 genetic variants; and MTAG can identify 1,934 genetic variants. From ACAT results, 6,060 genetic variants are identified by ACAT based on at least two other grouping of phenotypes, even if it identifies a smaller number of genetic variants. According to results from SHom and MTAG, tests based on the network modules detected from well-defined GPN identify more genetic variants than the tests based on the LDSC and the UK Biobank, but it identify less genetic variants than the tests based on the network modules detected from denser GPN.

2.4 Discussion

In this paper, we perform a comprehensive analysis to construct the bipartite genotype and phenotype networks (GPN), which can be a routine procedure in post-GWAS analyses. Owing to increasingly accessible GWAS summary statistics, the construction of GPN only requires the marginal association evidence between each genetic variant and each phenotype in GWAS summary data instead of accessing individual-level genotypes and phenotypes data. The denser representation of the bipartite GPN can be directly constructed by linking all genetic variants and phenotypes in GWAS summary datasets. Although a denser representation of bipartite GPN contains information on all pairwise associations between genetic variants and phenotypes, pruning the network makes biological sense and is computationally efficient[31]. The thresholding approach for pruning networks leads to stable network properties, but the threshold is significantly impacted by the size of a network (connectance). To address this, we propose to construct a well-defined GPN with a clear representation of genetic associations by comparing the network properties with a random network, including connectivity, centrality, and community structure. Our findings indicate that a well-defined network with an optimal threshold can preserve critical information on the associations between genetic variants and phenotypes.

Based on the construction of the denser and well-defined representation of bipartite GPN, we further propose two network topology annotations based on the degree centrality and the approximate betweenness centrality. Both of the annotations can be used to quantify the possibility of pleiotropy for genetic variants. We highlight one of our significant discoveries that link pleiotropy and disease heritability through the utilization of heritability enrichment analysis using the stratified LD score regression. We analyze 12 genetically correlated phenotypes to show that the genetic variants with high degree centrality and approximate betweenness centrality are enriched for disease heritability conditioning on known functional annotations from the baseline LD model. First, in analyses of the degree centrality based on the denser and the well-defined GPNs, we identify 10 phenotypes with significant heritability enrichment after using the LOPO approach. The significant enrichment indicates that the degree annotation can contribute more information to disease heritability if it is computed based on other highly genetically correlated phenotypes. We also observe that the denser GPN provides more information in the degree centrality as the degree centrality contains the strength of marginal association evidence. Second, we determine that network annotation based on the approximate

betweenness centrality calculated from the well-defined GPN is strongly enriched for disease heritability. However, the disease heritability of some phenotypes is fully explained by annotations from the baseline-LD model in the analysis of the approximate betweenness centrality calculated from the denser GPN.

Construction of the bipartite GPN also has important implications for the phenome-wide association studies (PheWAS). In particular, detecting the network modules of phenotypes from the constructed GPN is essential in understanding the global and local structures of associations between genetic variants and phenotypes, and in shedding light on association connections that may not be easily visible in the network topology. The detected network modules can be used as a priori grouping of phenotypes in PheWAS, then jointly testing of multiple phenotypes in each network module and one genetic variant can be performed to discover the cross-phenotype associations and pleiotropy. Significance thresholds for PheWAS are adjusted for multiple testing by applying the false discovery rate (FDR) control approach. First, we discover that the three multiple phenotype association tests (ACAT, SHom, and MTAG) applied in this study can well-control FDR as demonstrated by extensive simulation studies. Second, we analyze 633 EHR-derived phenotypes in the UK Biobank GWAS summary datasets. Based on the network modules detected from the denser representation of GPN, all three tests can identify more genetic variants associated with at least one network module (~6,000 genetic variants) compared with the tests based on the UK Biobank, LDSC, and well-defined GPN.

There still are some limitations to the work presented here. First, genetic effects can be heterogenous across phenotypes and studies based on different GWAS summary statistics[111,112] due to different sample sizes, genetic architectures, and quality of the genotyping and phenotyping data, et al. In our current analyses, we ignore the influence of different sample sizes for different GWAS summary statistics in the construction of GPN. However, larger sample sizes are typically associated with smaller standard errors and more precise effect size estimates, which can help to reduce bias and increase the stability of effect size estimates. To construct a GPN with stable evidence of the associations in the edges, we suggest that the sample sizes used to calculate the GWAS summary results in each study are sufficiently large (e.g., $N_k > 10,000$). Second, we use the marginal association between each genetic variant and each phenotype to define the edge of GPN. The challenge in validating our proposed construction of GPNs is that there is no source of genome-wide "ground truth". There may exist spurious associations between multiple genetic variants and a phenotype due to LD[3]. For example, a genetic variant in high LD with a true causal variant may be detected instead of the causal variant itself. However, several powerful fine-mapping and colocalization approaches have been developed to leverage information on LD to identify the putative causal variants in a specific genomic region[113-115], which provides a great opportunity to construct a more informative GPN for future studies. Third, we do not consider the functional relationships between genetic variants and phenotypes. Filtering candidate (functional) regions based on strength of powerful gene-based associations may reduce multiple testing burdens and consequently improve statistical power in the construction of GPN. For example, transcriptome-wide association studies can combine genetic and transcriptomic data in a specific tissue to

identify functional variants and genomic regions, which provide insights into biological pathways[116].

2.5 Data availability

GWAS summary statistics for 633 EHR-derived phenotypes with main ICD10 diagnoses can be found from Neale lab: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank.

GWAS summary statistics for 12 highly correlated phenotypes can be downloaded from the corresponding consortium websites reported in Zhang et al.[109].

LDSC: the command line tool for estimateing heritability and genetic correlation from GWAS summary statistiscs can be downloaded from https://github.com/bulik/ldsc[27].

Cytoscape: an open source software platform for visualizing complex networks which can be accessed via https://cytoscape.org/[71].

3 Chapter 3

Gene-based association tests using GWAS summary statistics and incorporating eQTL

Abstract

Although genome-wide association studies (GWAS) have been successfully applied to a variety of complex diseases and identified many genetic variants underlying complex diseases via single marker tests, there is still a considerable heritability of complex diseases that could not be explained by GWAS. One alternative approach to overcome the missing heritability caused by genetic heterogeneity is gene-based analysis, which considers the aggregate effects of multiple genetic variants in a single test. Another alternative approach is transcriptome-wide association study (TWAS). TWAS aggregates genomic information into functionally relevant units that map to genes and their expression. TWAS is not only powerful, but can also increase the interpretability in biological mechanisms of identified trait associated genes. In this study, we propose a powerful and computationally efficient gene-based association test, called Overall. Using extended Simes procedure, Overall aggregates information from three types of traditional gene-based association tests and also incorporates expression quantitative trait locus (eQTL) information into a gene-based association test using GWAS summary statistics. We show that after a small number of replications to estimate the correlation among the integrated gene-based tests, the *P* values of Overall can be calculated analytically. Simulation studies show that Overall can control type I error rates very well and has higher power than the tests that we compared with. We also apply Overall to two schizophrenia GWAS summary datasets and two lipids GWAS summary datasets. The results show that this newly developed method can identify more significant genes than other methods we compared with.

Keywords: extended Simes procedure; eQTL - derived weights; GWAS summary statistics; gene-based association study

3.1 Introduction

Although genome-wide association studies (GWAS) have successfully identified thousands of single nucleotide polymorphisms (SNPs) associated with a wide range of complex human traits[1,2], there is a common limitation in which GWAS focus on only a single genetic variant with a trait at a time. This limitation may limit the power to identify clinically or biologically significant genetic associations[117]. Furthermore, many genome-wide significant genetic variants are in linkage disequilibrium (LD). Different LD patterns can cause non-replicated results of the same variant in different populations[118,119]. Therefore, several powerful gene-based statistical association tests, in which the genetic information of all genetic variants in a gene is combined to obtain more informative results,

have been developed, such as the Burden Test (BT)[120], the Sequence Kernel Association Test (SKAT)[121], and the Optimized SKAT (SKATO)[122].

When individual-level genotype and phenotype data are not available, the traditional gene-based association tests, BT, SKAT, and SKATO, can be extended by using GWAS summary statistics[123]. Currently, there are many GWAS summary statistics available in public resources[124]. In GWAS summary statistics, the Z-scores of genetic variants in a gene are assumed to asymptotically follow a multivariate normal distribution with a correlation matrix among all genetic variants in a gene under the null hypothesis[125], where the correlation matrix can be estimated by LD among the genetic variants in the gene[116,126]. When individual-level data are not available, LD is usually estimated using external reference panels[127,128] (i.e., 1000 Genomes Project[129]). Due to the small sample size of reference panels used to estimate LD, statistical noise (i.e., inflated type I error rates or large numbers of false positives) often exists which needs to be accounted for[130,131]. One way to reduce the statistical noise is to correct the estimated LD by a regularization procedure[132]. In the regularization procedure, a statistical white Gaussian noise is added to the LD matrix which is estimated by a reference panel. After correcting the estimated LD by the regularization procedure, we can assume that, under the null hypothesis, the Z-scores from GWAS summary statistics asymptotically follow a multivariate normal distribution with the correlation matrix being the corrected LD matrix among the genetic variants in a gene.

To increase statistical power in identifying genes that are associated with complex diseases, PrediXcan[133] and transcriptome-wide association study[116,134] (TWAS) were developed by incorporating expression quantitative trait locus (eQTL) data into GWAS. As pointed out by Zhang et al.[128], PrediXcan and TWAS can be viewed as a simple weighted linear combination of genetic variants with an eQTL - derived weight. In fact, the genetic architecture of complex traits is rarely known in advance and is likely to vary from one region to another across the genome and from one trait to another[128]. Therefore, only considering one single eQTL - derived weight, such as in PrediXcan and TWAS, may lose statistical power in identifying significant genes. Zhang et al.[128] developed an omnibus test (OT) using Cauchy combination method to integrate association evidence obtained by BT, SKAT, and SKATO based on GWAS summary data with multiple eQTL-derived weights. They showed that OT using multiple eQTL - derived weights had some potential advantages.

Inspired by the advantage of OT, in this paper, we propose a more powerful and computationally efficient method, called Overall, to aggregate the information from three types of traditional gene-based association tests (BT, SKAT, SKATO) with multiple eQTL - derived weights using GWAS summary statistics. To combine information from the three gene-based association tests, the Overall method utilizes the extended Simes procedure[119,135]. To apply the Overall method, we first need to estimate the correlation matrix among the three gene-based association tests with eQTL - derived weights under the null hypothesis. We provide an estimation method using a replication procedure[136,137]. The replication procedure only needs to be performed once to obtain the correlation matrix for each gene. Then, the p-values of Overall can be analytically computed without using permutations. To calculate the p-values of the three types of gene-based association tests

(BT, SKAT, SKATO) using GWAS summary statistics with eQTL - derived weights, we use the "sumFREGAT" package in R[123]. Once we obtain the p-values of these three tests, the p-values of our proposed method can be easily calculated using its theoretical distribution. Extensive simulation studies show that Overall can control type I error rates well and has higher power than the comparison methods across most of the simulation settings. Similar to Zhang et al.[128], we apply our method to two schizophrenia (SCZ) and two lipids trait (HDL) GWAS summary data sets. Compared with OT and other tests, the proposed method can identify more significant genes. More interestingly, some significant genes reported by GWAS catalog are only identified by our proposed method.

3.2 Statistical Models and Methods

3.2.1 Statistical Models

Consider a set of M genetic variants in a gene. Let $\boldsymbol{Z} = \left(Z_1, \cdots, Z_M\right)^T$ be an $M \times 1$ vector of Z-scores of the M genetic variants. Note that the Z-scores is either directly provided by publicly available GWAS summary statistics or calculated from a GWAS individual-level genotype and phenotype data set. We are interested in testing the null hypothesis H_0 that none of the genetic variants in the gene is associated with a trait, whereas the alternative hypothesis is that at least one genetic variant in the gene is associated with a trait. Following Gusev et al.[116] and Yang et al.[138], we assume $\boldsymbol{Z} = \left(Z_1, \cdots, Z_M\right)^T \sim \text{MVN}\left(\boldsymbol{0}, \mathbf{R}\right)$ under the null hypothesis, where $\mathbf{R}$ is the correlation matrix among $\boldsymbol{Z}$, which can be estimated by LD among the genetic variants in the gene[116,126]. If individual-level data are not available, LD can be estimated using external reference panels (i.e., 1000 Genomes Project[129]). However, if the sample size of a reference panel is small, LD may not be estimated correctly so that it will induce statistical noise (i.e., inflated type I error rates or large numbers of false positives)[130,131]. One way to correct the estimated LD is to use a regularization procedure by adding a statistical white Gaussian noise[123,132]. Let $\mathbf{I}_M$ be an $M \times M$ identity matrix, and the corrected correlation matrix $\mathbf{U}$ can be defined as

$$\mathbf{U} = a\mathbf{R} + \left(1 - a\right)\mathbf{I}_M, \quad 0 \le a \le 1,$$

where a is a scalar tuning parameter which represents the coefficient of proportionality between the corrected correlation matrix $\mathbf{U}$ and the original $\mathbf{R}$ estimated using an external reference panel. The optimal tuning parameter a can be estimated by maximizing the log-likelihood function of the distribution of $\boldsymbol{Z} \sim \text{MVN}\left(\boldsymbol{0}, \mathbf{U}\right)$, that is,

$$\hat{a} = \arg\max_{a \in [0,1]} \left\{ \log\left(L\left(\boldsymbol{Z} : \boldsymbol{0}, \mathbf{U} \right) \right) \right\}.$$

Then the corrected correlation matrix $\hat{\mathbf{U}} = \hat{a}\mathbf{R} + \left(1 - \hat{a}\right)\mathbf{I}_M$. Therefore, under the null hypothesis, we consider $\boldsymbol{Z} = \left(Z_1, \cdots, Z_M\right)^T \sim \text{MVN}\left(\boldsymbol{0}, \hat{\mathbf{U}}\right)$.

Suppose that there are a total of K different eQTL - derived weights from gene expression data (i.e., Genotype-Tissue Expression (GTEx; https://gtexportal.org/home/)),

denoted as $\hat{\mathbf{W}}_k = \text{diag}\left(\hat{W}_1^k, \cdots, \hat{W}_M^k\right)$ for $k = 0,1,\cdots,K$, where $\hat{\mathbf{W}}_0 = \text{diag}(1,\cdots,1)$ represents a status without using any weight. In order to avoid the influence of the scale among genetic variants within each weight, we first standardize the eQTL - derived weights $\mathbf{W}_k$ as $W_m^k = \hat{W}_m^k \Big/ \sum_{m=1}^{M} \left|\hat{W}_m^k\right|$ for $m = 1,\cdots,M$. Based on the k^{th} standardized weight $\mathbf{W}_k$, the weighted Z-score $\mathbf{W}_k \boldsymbol{Z}$ follows a multivariate normal distribution. That is,

$$\mathbf{W}_k \boldsymbol{Z} \sim \text{MVN}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_k\right) \text{ and } \hat{\boldsymbol{\Sigma}}_k = \mathbf{W}_k \hat{\mathbf{U}} \mathbf{W}_k.$$

We extend the three types of gene-based association tests, BT[120], SKAT[121], and SKATO[122], to incorporate the eQTL - derived weights based on GWAS summary statistics[123,139]. For the k^{th} eQTL - derived weight, the three gene-based test statistics can be written as

$$Q_{BT}^k = \left(\boldsymbol{Z}^T \mathbf{W}_k \boldsymbol{1}_M\right)^2,$$

$$Q_{SKAT}^k = \left(\mathbf{W}_k \boldsymbol{Z}\right)^T \mathbf{W}_k \boldsymbol{Z},$$

$$Q_{SKATO}^k = \min_{\rho \in [0,1]} \left\{(1-\rho) Q_{SKAT}^k + \rho Q_{BT}^k\right\},$$

where $\boldsymbol{1}_M$ is an $M \times 1$ vector with elements of all 1s. Under the null hypothesis, Q_{BT}^k follows a χ^2 distribution with 1 degree of freedom; Q_{SKAT}^k follows a weighted sum of χ^2 distributions with 1 degree of freedom; and Q_{SKATO}^k follows a mixture of χ^2 distribution[122]. The p-values of these three test statistics can be easily calculated using the "sumFREGAT" package in R[123].

3.2.2 Overall Method

To aggregate information from these three gene-based association tests with multiple eQTL - derived weights, we develop a novel method, called Overall, which utilizes the extended Simes procedure[119,135]. Let $p_{BT}^k, p_{SKAT}^k, p_{SKATO}^k$ be the p-values of BT, SKAT, SKATO with k^{th} eQTL - derived weight, $k = 0,1,\cdots,K$, respectively, where $k = 0$ denotes a status without using any weight. Thus, there are a total of $L = 3(K+1)$ p-values from three gene-based tests with different weights. Let $\left(p_{(1)}, \cdots, p_{(L)}\right)$ be a sequence of the ascending p-values with $p_{(1)} = \min_{k=0,\cdots,K} \left\{p_{BT}^k, p_{SKAT}^k, p_{SKATO}^k\right\}$ and $p_{(L)} = \max_{k=0,\cdots,K} \left\{p_{BT}^k, p_{SKAT}^k, p_{SKATO}^k\right\}$. Overall combines these L p-values using the extended Simes procedure[119,135], and the p-value of Overall is defined as

$$p_{overall} = \underset{l=1,\cdots,L}{Min} \left\{\frac{m_e p_{(l)}}{m_{e(l)}}\right\},$$

where m_e is the effective number of p-values among the L gene-based association tests with multiple weights, $p_{(l)}$ is the l^{th} element of the ascending p-values, and $m_{e(l)}$ is the

effective number of p-values among the top l association tests. We use a more robust measure to obtain the effective numbers m_e and $m_{e(l)}$, which was proposed by Li et al.[119]. The values of $m_{e(l)}$ and m_e can be estimated as

$$m_{e(l)} = l - \sum_{i=1}^{l} \left[\left(\lambda_i - 1 \right) I\left(\lambda_i > 1 \right) \right] \text{ and } m_e = m_{e(L)},$$

where λ_i denotes the i^{th} eigenvalue of the correlation matrix $\boldsymbol{\Omega}$ of p-values from L association tests with multiple weights (the estimation of $\boldsymbol{\Omega}$ will be discussed in the next section), $I(\bullet)$ is an indicator function. If the L association tests are independent, all eigenvalues λ_i equal 1, and $m_{e(l)} = l$ for $l = 1, \cdots, L$; if the L association tests are perfectly dependent, then $\lambda_1 = l$ which is the number of tests used to calculate $m_{e(l)}$ and the other eigenvalues equal 0. In this case, $m_{e(l)} = l - (l-1) = 1$ for $l = 1, \cdots, L$.

The R codes and a sample data set for the implementation of Overall are available at GitHub https://github.com/xueweic/Overall.

3.2.3 Estimation of Ω under the null hypothesis

To apply our proposed method, we need to estimate the correlation matrix of p-values $\boldsymbol{\Omega}$ under the null hypothesis. Since the exact correlations among all L gene-based association tests are unknown, we perform the estimation procedure with B replications. For each replicate b, $b = 1, \cdots, B$, we implement the following two steps:

Step 1: We first generate a new Z-score vector $\boldsymbol{Z}^{null}$ under the null hypothesis. That is, $\boldsymbol{Z}^{null}$ follows a multivariate normal distribution with mean $\boldsymbol{0}$ and variance-covariance matrix $\mathbf{R}$, where $\mathbf{R}$ can be estimated by LD among the genetic variants in a gene using external reference panels (i.e., 1000 Genomes Project).

Step 2: We use the regularization procedure to obtain the corrected correlation matrix of Z-scores $\hat{\mathbf{U}}$. Then, we calculate $Q_{BT}^{k(b)}, Q_{SKAT}^{k(b)}, Q_{SKATO}^{k(b)}$ and the corresponding p-values $p_{BT}^{k(b)}, p_{SKAT}^{k(b)}, p_{SKATO}^{k(b)}$ using the simulated $\boldsymbol{Z}^{null}$ for $k = 0, 1, \cdots, K$. The distributions of $Q_{BT}^{k(b)}, Q_{SKAT}^{k(b)}, Q_{SKATO}^{k(b)}$ depend on the corrected correlation matrix $\hat{\mathbf{U}}$, and the standardized eQTL - derived weights $\mathbf{W}_k$ for $k = 0, 1, \cdots, K$.

To estimate the correlation matrix of p-values $\boldsymbol{\Omega}$ used in the Overall method, we use the sample correlation matrix of the p-values obtained from the replications. We denote the sample correlation matrix of p-values as $\hat{\boldsymbol{\Omega}}$. For example, $\hat{\Omega}_{12}$ is the (1,2)-element of $\hat{\boldsymbol{\Omega}}$ which is the estimated correlation between BT and SKAT without using any weight. If we let $\boldsymbol{p}_{BT}^{0} = \left(p_{BT}^{0(1)}, \cdots, p_{BT}^{0(B)} \right)^T$ be a $B \times 1$ vector of the p-values of BT without using any weight and $\boldsymbol{p}_{SKAT}^{0} = \left(p_{SKAT}^{0(1)}, \cdots, p_{SKAT}^{0(B)} \right)^T$ be a $B \times 1$ vector of the p-values of SKAT without using any weight obtained from the replications, then the sample correlation of p-values

between these two tests is defined as $\hat{\Omega}_{12} = \text{cor}\left(\boldsymbol{p}_{BT}^{0}, \boldsymbol{p}_{SKAT}^{0}\right)$, where $\text{cor}(\bullet)$ is the sample correlation.

The estimation procedure to estimate $\boldsymbol{\Omega}$ is independent of our proposed method, therefore we only need to perform this procedure once for each gene. After we estimate $\boldsymbol{\Omega}$, the p-value of Overall can be computed analytically without using permutations.

3.3 Simulation Studies

3.3.1 Materials and Comparison Methods

In our studies, we use four data sets to obtain the eQTL - derived weights downloaded from the website (http://gusevlab.org/projects/fusion/#reference-functionaldata). The resources to obtain the four eQTL - derived weights are listed in **Table 3.1**. For each eQTL data set, we use the weights estimated by the Best Linear Unbiased Prediction (BLUP)[140].

Table 3.1. Resources of the four eQTL - derived weights used in the simulation studies.

Study	Tissue	# of Samples	Reference
NTR	Peripheral blood	1247	Wright et al.[141]
YFS	Whole blood	1264	Gusev et al.[116]
METSIM	Adipose	563	Gusev et al.[116]
CMC	Brain	452	Gusev et al.[116]

We compare our proposed method with three existing methods, OT[128] , S-PrediXcan[142], and S-TWAS[116]. These three methods are all based on GWAS summary statistics and incorporate eQTL-derived weights. Here, we briefly introduce these methods.

OT: For a total of K different eQTL - derived weights and the three gene-based association tests (BT, SKAT, SKATO), OT aggregates information from different weights and tests by using the Cauchy combination method[143]. The test statistic of OT is defined as $Q_{OT} = \frac{1}{3(K+1)}\sum_{k=0}^{K}\left[\tan\left\{\left(0.5-p_{BT}^{k}\right)\pi\right\}+\tan\left\{\left(0.5-p_{SKAT}^{k}\right)\pi\right\}+\tan\left\{\left(0.5-p_{SKATO}^{k}\right)\pi\right\}\right]$ and the corresponding p-value can be approximated by $p_{OT} = 1/2 - \arctan\left(Q_{OT}\right)/\pi$.

S-PrediXcan: For a given eQTL-derived weight, provided by a matrix $\mathbf{W}_k = \text{diag}\left(W_1^k, \cdots, W_M^k\right)$, the test statistic of S-PrediXcan is $Z_{S\text{-}PrediXcan}^{k} = \sum_m W_m^k \hat{\sigma}_m Z_m / \hat{\sigma}$, where $\hat{\sigma}_m$ is the estimated standard deviation of the m^{th} SNP in a gene and $\hat{\sigma}$ is the estimated standard deviation of the predicted expression of a gene. The p-value of S-PrediXcan can be computed as $p_{S\text{-}PrediXcan}^{k} = 2\Phi\left(-\left|Z_{S\text{-}PrediXcan}^{k}\right|\right)$, where $\Phi(\bullet)$ is the standard normal CDF function.

S-TWAS: For a given eQTL-derived weight, provided by a vector $\boldsymbol{w}_k = \left(W_1^k, \cdots, W_M^k\right)^T$, the test statistic of S-TWAS is defined as $Z_{S\text{-}TWAS}^{k} = \frac{\boldsymbol{w}_k^T \cdot \boldsymbol{Z}}{\sqrt{\boldsymbol{w}_k^T \cdot \mathbf{R} \cdot \boldsymbol{w}_k}}$, where $\mathbf{R}$ is the estimated

LD structure among the genetic variants in a gene and the corresponding p-value can be calculated by $p_{S\text{-}TWAS}^{k} = 2\Phi\left(-\left|Z_{S\text{-}TWAS}^{k}\right|\right)$.

3.3.2 The Number of Replications Needed in Estimation of $\boldsymbol{\Omega}$

To apply our proposed method, we first need to estimate the correlation matrix of p-values, $\boldsymbol{\Omega}$, under the null hypothesis for each gene. Following the estimation procedure introduced in the method section, we generate Z-scores instead of generating individual-level genotype and phenotype data. To determine the number of replications needed in the estimation of $\boldsymbol{\Omega}$, we consider 18 genes that contain different numbers of SNPs and have different LD structures. **Table C.1** gives a summary of these 18 genes. We can see from **Table C.1**, the number of SNPs in a gene is ranging from 23 to 359 and the average per-SNP LD score in a gene is ranging from 12.72 to 170.85. We simulate a Z-score vector from a multivariate normal distribution with mean $\boldsymbol{0}$ and variance-covariance matrix $\mathbf{R}$, $\boldsymbol{Z} \sim \text{MVN}(\boldsymbol{0}, \mathbf{R})$, where $\mathbf{R}$ is the LD matrix of each gene which can be estimated using the 1000 Genomes Project (unrelated Europeans in 1000 Genomes in Phase 3; ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/). First, we use $B = 10^4$ replications to estimate $\boldsymbol{\Omega}$ under the null hypothesis, where the estimated matrix is denoted by $\hat{\boldsymbol{\Omega}}$. Then, we denote $\hat{\boldsymbol{\Omega}}^0$ as the correlation matrix of p-values by using B_0 replications. We vary the value of B_0 from 16 to 5,000, and test the null hypothesis that the two correlation matrices, $\hat{\boldsymbol{\Omega}}^0$ and $\hat{\boldsymbol{\Omega}}$, are the same by using "lavaan" package (https://CRAN.R-project.org/package=lavaan)[144]. **Figure C.1** shows that the p-values for the hypothesis testing in each gene are greater than 0.05 after $B_0 = 1{,}000$ replications for all of the 18 genes. Therefore, we recommend using 1,000 replications to obtain $\hat{\boldsymbol{\Omega}}$ for each gene under the null hypothesis. Consequently, 1,000 replications are used in the following sessions to evaluate the type I error rates and powers of Overall.

3.3.3 Type I error rates

To evaluate if our proposed method can control type I error rates, we perform simulations based on the aforementioned 18 genes. For each of the 18 genes, we generate Z-score vectors under the null hypothesis, $\boldsymbol{Z} \sim \text{MVN}(\boldsymbol{0}, \mathbf{R})$, where $\mathbf{R}$ is the LD matrix of the gene estimated using the 1000 Genomes project. Then, we use the regularization procedure to obtain the corrected correlation matrix of Z-scores $\hat{\mathbf{U}}$, and calculate the three types of gene-based association tests, BT, SKAT, and SKAT-O, with or without the four eQTL-derived weights (NTR, YFS, METSIM, CMC) based on the corrected correlation matrix $\hat{\mathbf{U}}$. Finally, we apply our proposed method to combine the p-values using the estimated correlation matrix of p-values, $\hat{\boldsymbol{\Omega}}$, with 1,000 replications.

We generate simulated data to mimic real lipids data which we will use in real data analysis section. Gene *AGTRAP* is associated with lipids trait HDL[15], There are a total of 23 genetic variants in gene *AGTRAP*. The LD block structure of these 23 genetic variants is shown in **Figure C.**2. **Figure C.3** shows the estimated correlation matrix $\hat{\boldsymbol{\Omega}}$ for this

gene. We use 10^7 replications to evaluate type I error rates of Overall for gene *AGTRAP* at 5×10^{-2}, 1×10^{-2}, 1×10^{-3}, 1×10^{-4},1×10^{-5}, and 1.75×10^{-6} significance levels. With 10^7 replications, a Bonferroni corrected significance level of 1.75×10^{-6} can be reached to obtain the empirical type I error rates (i.e., for 28,625 genes in the real data analysis section, the Bonferroni corrected significance level is $0.05/28625=1.75\times10^{-6}$ at 5% significance level). We further evaluate type I rates based on the other 17 genes. To save computational time, we use 2×10^5 replications to evaluate type I error rates of Overall for the 17 genes at significance levels of 1×10^{-2}, 1×10^{-3}, and 1×10^{-4}. **Table 3.2** and **Tables C.2** show the estimated type I error rates of Overall under various nominal significance levels for gene *AGTRAP and the other 17 genes*, respectively. From these tables, we can see that our proposed method can control type I error rates very well at different significant levels.

Table 3.2. Estimated type I error rates at different significance levels with 10^7 replications. The subscript denotes BT, SKAT, and SKATO using eQTL - derived weights; CMC, METSIM, NTR, and YFS indicate the resources to obtain the eQTL - derived weights. 0 indicates the methods without using eQTL – derived weights.

$\alpha-$ Level	5×10^{-2}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}	1.75×10^{-6}
BT_0	5.03×10^{-2}	1.06×10^{-2}	1.00×10^{-3}	1.01×10^{-4}	9.76×10^{-6}	1.84×10^{-6}
$SKAT_0$	5.24×10^{-2}	1.07×10^{-2}	1.01×10^{-3}	1.00×10^{-4}	1.04×10^{-5}	1.80×10^{-6}
$SKATO_0$	4.58×10^{-2}	9.57×10^{-3}	1.02×10^{-3}	1.04×10^{-4}	9.72×10^{-6}	1.46×10^{-6}
BT_{CMC}	5.17×10^{-2}	1.04×10^{-2}	1.01×10^{-3}	9.82×10^{-5}	9.58×10^{-6}	1.72×10^{-6}
$SKAT_{CMC}$	5.08×10^{-2}	9.89×10^{-3}	9.71×10^{-4}	9.75×10^{-5}	9.48×10^{-6}	1.66×10^{-6}
$SKATO_{CMC}$	5.16×10^{-2}	1.09×10^{-2}	1.17×10^{-3}	1.21×10^{-4}	1.22×10^{-5}	2.14×10^{-6}
BT_{METSIM}	5.02×10^{-2}	1.03×10^{-2}	1.02×10^{-3}	1.01×10^{-4}	9.86×10^{-6}	1.66×10^{-6}
$SKAT_{METSIM}$	5.30×10^{-2}	1.08×10^{-2}	1.02×10^{-3}	9.91×10^{-5}	1.00×10^{-5}	2.12×10^{-6}
$SKATO_{METSIM}$	4.84×10^{-2}	1.05×10^{-2}	1.11×10^{-3}	1.09×10^{-4}	1.06×10^{-5}	1.84×10^{-6}
BT_{NTR}	5.02×10^{-2}	1.06×10^{-2}	1.00×10^{-3}	9.93×10^{-5}	1.01×10^{-5}	1.76×10^{-6}
$SKAT_{NTR}$	5.09×10^{-2}	1.03×10^{-2}	9.98×10^{-4}	1.00×10^{-4}	1.01×10^{-5}	2.00×10^{-6}
$SKATO_{NTR}$	5.08×10^{-2}	1.18×10^{-2}	1.34×10^{-3}	1.45×10^{-4}	1.52×10^{-5}	2.92×10^{-6}
BT_{YFS}	5.10×10^{-2}	1.02×10^{-2}	9.95×10^{-4}	9.95×10^{-5}	1.05×10^{-5}	2.10×10^{-6}
$SKAT_{YFS}$	4.98×10^{-2}	1.03×10^{-2}	9.97×10^{-4}	1.01×10^{-4}	1.02×10^{-5}	2.06×10^{-6}
$SKATO_{YFS}$	5.58×10^{-2}	1.32×10^{-2}	1.43×10^{-3}	1.55×10^{-4}	1.69×10^{-5}	3.50×10^{-6}
Overall	4.67×10^{-2}	1.01×10^{-2}	1.12×10^{-3}	1.14×10^{-4}	1.24×10^{-5}	2.44×10^{-6}

3.3.4 Power Comparison

To evaluate the performance of the Overall method, we use several simulations to compare the power of Overall with the power of OT, S-PrediXcan, S-TWAS, and three types of gene-based association tests with and without eQTL - derived weights. We use BEST to represent the test with the maximum power among the three traditional gene-based association tests with and without an eQTL - derived weight, S-TWAS.B and S-PrediXcan.B to represent the maximum power of S-TWAS and S-PrediXcan with each of the eQTL – derived weights, respectively. Following the simulation settings in Nagpal et al.[145] and Zhang et al.[128], we generate individual-level genotypes, phenotypes, and different gene expression levels using the following steps:

(1) The genotype data are generated using the haplotypes of a gene obtained from the 1000 Genomes Project reference panel. To generate the genotype of an individual, $\mathbf{X}_g$, we select two haplotypes according to the haplotype frequencies from the haplotype pool and then remove genetic variants with MAF<0.05.

(2) We consider K different weights derived from gene expression data which can be estimated using BLUP. To generate a vector of weights, $\boldsymbol{w}_k$, for the k^{th} gene expression level, we randomly select causal variants according to the proportion of causal variants, p_{causal}. Then, the effect sizes for the k^{th} gene expression levels and M_{causal} causal variants can be generated from a standard normal distribution, $w_{mk} \sim N(0,1)$ for $m = 1, \cdots, M_{causal}$, where $M_{causal} = M \times p_{causal}$; otherwise, $w_{mk} = 0$. After we rescaled the weights to ensure the targeted expression heritability h_e^2, we generate the k^{th} gene expression level by $\boldsymbol{E}_k = \mathbf{X}_g \boldsymbol{w}_k + \boldsymbol{\varepsilon}_e$ with each element of random error $\boldsymbol{\varepsilon}_e$ follows $N(0, 1-h_e^2)$.

(3) Let $\mathbf{E} = (\boldsymbol{E}_1, \cdots, \boldsymbol{E}_K)$ be the matrix of gene expression levels. Phenotypes are generated by using a formula $\boldsymbol{Y} = \mathbf{E}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_p$ with each element of random error $\boldsymbol{\varepsilon}_p$ follows $N(0, 1-h_p^2)$, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_K)^T$ is a vector of genetic effect sizes which can be assigned based on the phenotypic heritability h_p^2.

(4) The Z-score vector is estimated from individual-level genotype and phenotype data using beta coefficient and its standard division estimated based on the ordinary least squares method in linear regression.

In our simulation studies for power comparison, we consider two genes, AGTRAP and *C3orf22,* from the 18 genes used in the type I error evaluation and $K = 4$ and $K = 20$ eQTL - derived weights. *AGTRAP* contains 458 haplotypes for 23 genetic variants (11 common variants and 12 rare variants; MAF ranging from 0 to 0.39775); *C3orf22* contains 295 haplotypes for 42 variants (18 common variants and 24 rare variants; MAF ranging from 0 to 0.43558). **Figure C.2** shows the LD block structure of the 23 genetic variants at *AGTRAP* and the 42 genetic variants at *C3orf22*. We vary the proportion of causal variants with $p_{causal} = (0.2, 0.3, 0.4, 0.5)$ for *AGTRAP* and $p_{causal} = (0.1, 0.2, 0.3, 0.4)$ for *C3orf22*. We also consider two different directions of genetic effects: $\beta_1 = \cdots = \beta_K$ (Scenario 1: Uni-directional effects) and $\beta_1 = \cdots = \beta_{K/2} = -\beta_{K/2+1} = \cdots = -\beta_K$ (Scenario 2: Bi-directional effects). For each simulation scenario, we vary the proportion of gene expression heritability and the phenotypic heritability with different values of h_e^2 and h_p^2. We consider the sample size to be 2,000 (unless it is specified) and the power is calculated as the proportion of 1,000 replications with p-value $< 1.75 \times 10^{-6}$.

Figure 3.1 (**Figure C.4**) shows the power comparisons based on gene *AGTRAP* (and *C3orf22*) with $K = 4$ under the Uni-directional effects ($\beta_1 = \beta_2 = \beta_3 = \beta_4$) with

different p_{causal}. We consider two settings here. First, we vary phenotypic heritability h_p^2 with a fixed expression heritability $h_e^2 = 0.2$ (**Figure 3.1(a)** and **Figure C.4(a)**). Second, we vary the expression heritability h_e^2 with a fixed phenotypic heritability $h_p^2 = 0.2$ (**Figure 3.1(b)** and **Figure C.4(b)**). **Figure 3.2** (**Figure C.5**) shows power comparisons based on gene *AGTRAP* (and *C3orf22*) under the Bi-directional effects ($\beta_1 = \beta_2 = -\beta_3 = -\beta_4$) with different p_{causal} for $K = 4$. We also consider two simulation settings, power against the phenotypic heritability h_p^2 with a fixed expression heritability $h_e^2 = 0.2$ and power against the expression heritability h_e^2 with a fixed phenotypic heritability $h_p^2 = 0.2$. The pattern of the power in **Figure 3.2** (**Figure C.5**) is similar to what we observe in **Figure 3.1** (**Figure C.4**). These figures show that (1) Overall and OT perform uniformly better than BEST, S-TWAS.B, and S-PrediXcan.B. We can see that Overall and OT boost power significantly due to integrating association evidence by different traditional tests and multiple eQTL – derived weights. Overall is slightly more powerful than OT in all of the scenarios. (2) Among BEST, S-TWAS.B, and S-PrediXcan.B, BEST are more powerful than S-TWAS.B and S-PrediXcan.B in all of the scenarios for gene *C3orf22*; For gene *AGTRAP*, S-TWAS.B and S-PrediXcan.B perform better than BEST when the proportion of causal variants in a gene is small ($p_{causal} = (0.2, 0.3)$); otherwise, BEST performs better than S-TWAS.B and S-PrediXcan.B.

To evaluate if Overall and OT that integrate different types of association tests and multiple eQTL – derived weights are robust for more eQTL studies, we also consider 20 ($K = 20$) eQTL - derived weights under Uni-directional effect and Bi-directional effect models on gene *C3orf22* with settings similar to the settings in **Figures C.4 and C.5**. After integrating $L = 3(K+1) = 63$ traditional gene-based association tests, we observe that the patterns of the power for $K = 20$ are similar to that in **Figures C.4 and C.5** with $K = 4$, and the power gain of Overall and OT is higher than that of the tests only consider one eQTL – derived weight, such as BEST, S-PrediXcan.B, and S-TWAS.B (**Figure C.6**).

Furthermore, we consider simulation settings with noise to the eQTL. We consider simulation settings by adding less noise to the eQTL from the most relevant tissues and more noise to those from the less relevant tissues. For the Uni-direction scenario, we consider the first study being the most relevant tissue, where $\beta_1 = \beta_0 + N\left(0, 0.1h_p^2\right)$ and $\beta_2 = \beta_3 = \beta_4 = \beta_0 + N\left(0, 0.5h_p^2\right)$; $\beta_0 = \sqrt{h_p^2 / K}$ depends on the phenotypic heritability h_p^2. For the Bi-direction scenario, we consider 1st and 3rd studies being the most relevant tissues that have opposite effect directions, where $\beta_1 = -\beta_0 + N\left(0, 0.1h_p^2\right)$, $\beta_3 = \beta_0 + N\left(0, 0.1h_p^2\right)$ and $\beta_2 = -\beta_0 + N\left(0, 0.5h_p^2\right)$, $\beta_4 = \beta_0 + N\left(0, 0.5h_p^2\right)$. Other parameter settings are the same as these in **Figures C.4 and C.5**. The power comparison results are shown in **Figures C.7 and C.8**. From these figures, we find that the patterns of the power in **Figures C.7 and C.8** are very similar to those in **Figures C.4 and C.5**.

(a) $h_e^2 = 0.2$ and $h_p^2 = 0.2, 0.4, 0.6, 0.8$ (b) $h_p^2 = 0.2$ and $h_e^2 = 0.2, 0.4, 0.6, 0.8$

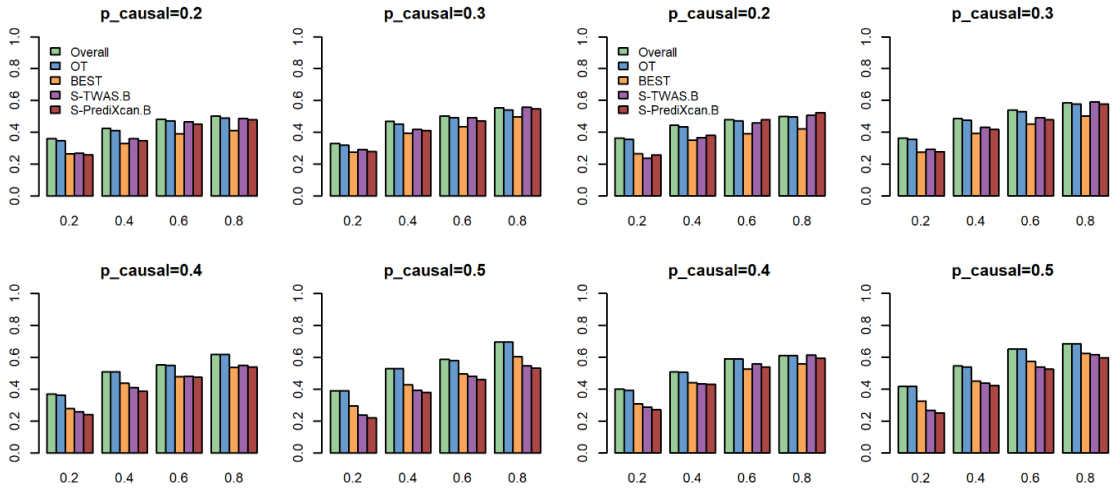


Figure 3.1. Power comparisons of gene-based association tests at 1.75×10^{-6} significance level under Uni-directional effects ($\beta_1 = \beta_2 = \beta_3 = \beta_4$) with $p_{causal} = (0.2, 0.3, 0.4, 0.5)$ based on gene *AGTRAP*. (a) Estimated power against phenotypic heritability h_p^2 with fixed expression heritability $h_e^2 = 0.2$; (b) Estimated power against expression heritability h_e^2 with fixed phenotypic heritability $h_p^2 = 0.2$.

(a) $h_e^2 = 0.2$ and $h_p^2 = 0.2, 0.4, 0.6, 0.8$ (b) $h_p^2 = 0.2$ and $h_e^2 = 0.2, 0.4, 0.6, 0.8$

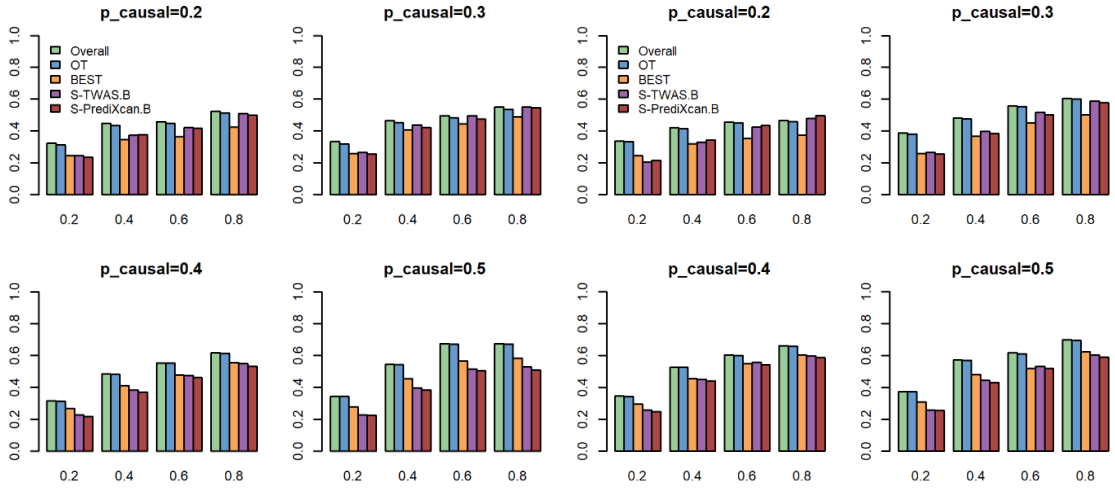


Figure 3.2. Power comparisons of gene-based association tests at 1.75×10^{-6} significance level under Bi-directional effects ($\beta_1 = \beta_2 = -\beta_3 = -\beta_4$) with $p_{causal} = (0.2, 0.3, 0.4, 0.5)$ based on gene *AGTRAP*. (a) Estimated power against phenotypic heritability h_p^2 with expression heritability $h_e^2 = 0.2$; (b) Estimated power against expression heritability h_e^2 with phenotypic heritability $h_p^2 = 0.2$.

In all of the previous power comparisons, we use a sample size of 2,000. We also consider simulation settings as those in **Figures C.7 and C.8**, but with a large sample size of 100,000. **Figure C.9** shows the results of power comparisons. We can see from this figure, all powers are increased with this larger sample size, but the patterns of the power are very similar to those in **Figures C.7 and C.8**.

To remove noise in LD matrix computed from a reference sample, we shrink the observed LD matrix toward an identity matrix with the shrinkage parameter estimated by maximum likelihood. To evaluate how well this regulation process performs, we compare the powers of three traditional gene-based association tests with and without eQTL – derived weights, OT, and Overall based on corrected and uncorrected LD structure. We use the same simulation settings as those in **Figures C.7 and C.8**. **Figure C.10** shows the power comparison results based on gene *C3orf22* under Uni-directional effects and Bi-directional effects with noise to eQTL. We can see that the powers of these tests based on corrected LD structure perform better than those based on uncorrected LD structure in most of the settings.

3.4 Real Data Analysis

To evaluate the performance of our proposed method, we apply Overall, OT, the three traditional tests with and without eQTL - derived weights, S-PrediXcan, and S-TWAS to the GWAS summary statistics data sets used in Zhang et al.[128]: two SCZ GWAS summary data sets and two lipid GWAS summary data sets. We estimate the LD between genetic variants using the 1000 Genomes Project reference panel[129], and obtain the corrected matrix of Z-score after the regularization procedure. We consider four eQTL - derived weights estimated by the BLUP method using the resources listed in **Table 3.1** (NTR, YFS, METSIM, CMC).

3.4.1 Application to the SCZ GWAS summary data

We consider two SCZ GWAS summary data sets, SCZ1 and SCZ2, which can be downloaded from the Psychiatric Genomics Consortium website (https://www.med.unc.edu/pgc/results-and-downloads/)[146]. SCZ1 is a meta-analysis of SCZ GWAS data set with 13,833 cases and 18,310 controls. SCZ2 is a more recent and larger SCZ GWAS summary data set with 36,989 cases and 113,075 controls for partial validation[147]. In our real data analysis, we define a gene to include all of the SNPs from 20 kb upstream to 20 kb downstream of the gene and test the association between each gene and the trait. We consider all genes according to the GENCODE version 35 (GRCh37) human comprehensive gene annotation list which can be downloaded from the GENCODE website (https://www.gencodegenes.org/human/release_35lift37.html).

To make fair comparisons among all these weighted tests, the genetic variants are removed if there is at least one weight missing in the four eQTL - derived weights. After pruning, there are 26,575 genes in SCZ1 and 17,823 genes in SCZ2 left in our final analyses. Therefore, the Bonferroni corrected significance level for gene-based association analysis is defined as 0.05 divided by the number of genes. First, we apply BT, SKAT, and SKATO with and without an eQTL - derived weight, OT, Overall, S-PrediXcan, and S-TWAS to the SCZ1 and SCZ2 data sets. **Table 3.3** (SCZ1 and SCZ2) shows the number of genes

identified by each method for the SCZ data sets, respectively. As we can see in **Table 3.3**, Overall identifies more genes than all of the other methods for two SCZ GWAS summary data sets. Among the three types of gene-based association tests, BT, SKAT, and SKATO, with or without different eQTL – derived weights, $SKATO_0$ identifies the greatest number of genes. $S\text{-}TWAS_{YFS}$ and $S\text{-}PrediXcan_{YFS}$ identify the greatest number of genes compared with S-TWAS and PrediXcan based on the other three eQTL – derived weights, respectively. Therefore, in **Figure 3.3**, we only show the number of genes identified by Overall, OT, $SKATO_0$, $S\text{-}PrediXcan_{YFS}$, and $S\text{-}TWAS_{YFS}$. The number below each method indicates the total number of genes identified by the corresponding method. From **Figure 3.3**, we can see that Overall identifies all of the genes identified by OT for SCZ1; for SCZ2, there are two genes identified by OT but failed to be identified by Overall; there are 66 and 24 genes identified only by Overall for SCZ1 data and SCZ2, respectively.

We further investigate the 90 genes identified only by Overall for the SCZ data sets by searching the GWAS catalog (https://www.ebi.ac.uk/gwas/). Among the 66 genes for the SCZ1 data set, there are six genes reported in the GWAS catalog; among the 24 genes for the SCZ2 data set, there are six genes reported in the GWAS catalog (**Table 3.4**). We also use these two SCZ GWAS data sets for partial validation. **Table 3.3** shows that there are 45 overlapping genes identified by Overall using SCZ1 and SCZ2 data sets and only 17 overlapping genes identified by OT using both SCZ1 and SCZ2 data sets. Furthermore, we search for genome-wide significant SNPs ($p<5\times10^{-8}$) from the two SCZ GWAS summary data sets and consider the genes covering at least one genome-wide significant SNP from 20 kb upstream to 20 kb downstream of the gene. There are 63 genome-wide significant genes for SCZ1, and 2422 genome-wide significant genes in SCZ2. **Table 3.3** ($GWAS_{SCZ1}$ and $GWAS_{SCZ2}$) summarizes the numbers of genome-wide significant genes that are identified by each method for the two SCZ data sets. Among the 63 genome-wide significant genes for the SCZ1 data set, Overall identifies the greatest number of genes, followed by $SKAT_0$ and $SKATO_0$; OT, $S\text{-}PrediXcan_{NTR}$ and $S\text{-}TWAS_{NTR}$ only identify 6 genes. Meanwhile, among 2422 genome-wide significant genes for SCZ2, Overall identifies 167 genes; OT identifies 166 genes; SKATO and $SKATO_0$ identify 153 genes; $S\text{-}TWAS_{YFS}$ and $S\text{-}PrediXcan_{YFS}$ only identify 58 and 72 genes respectively.

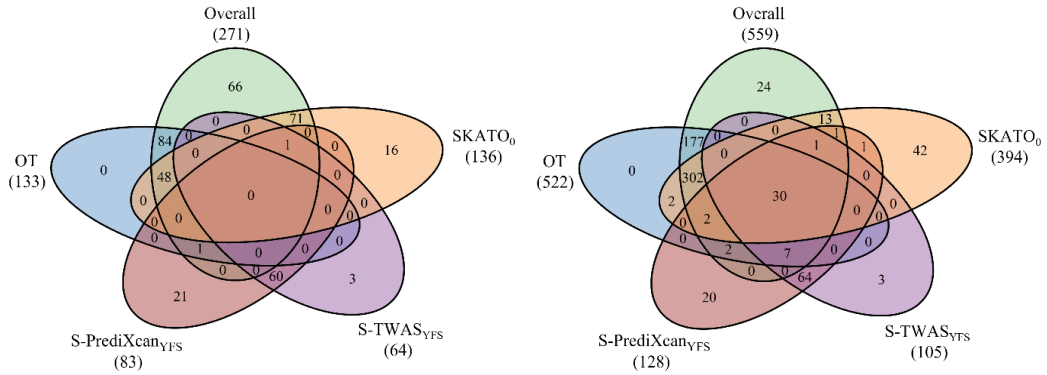


Figure 3.3. Venn diagram of the number of genes identified by Overall, OT, $SKATO_0$, $S\text{-}PrediXcan_{YFS}$, and $S\text{-}TWAS_{YFS}$ for SCZ1 data (left) and SCZ2 data (right). The number below each of the methods indicates the total number of significant genes identified by the corresponding method.

Table 3.3. The numbers of genes identified by each method for the two SCZ data sets. The subscript denotes BT, SKAT, and SKATO using eQTL - derived weights; CMC, METSIM, NTR, and YFS indicate the resources to obtain the eQTL - derived weights. 0 indicates the methods without using any weights.

	SCZ1	**SCZ2**	**$SCZ_{overlap}$**	**$GWAS_{SCZ1}$**	**$GWAS_{SCZ2}$**
BT_0	97	166	7	1	38
$SKAT_0$	47	305	20	15	153
$SKATO_0$	136	394	27	15	153
BT_{CMC}	44	137	2	1	56
$SKAT_{CMC}$	12	225	6	1	134
$SKATO_{CMC}$	30	263	2	1	130
BT_{METSIM}	44	136	5	1	48
$SKAT_{METSIM}$	23	223	9	4	132
$SKATO_{METSIM}$	31	205	3	0	100
BT_{NTR}	48	119	7	6	48
$SKAT_{NTR}$	27	230	9	8	141
$SKATO_{NTR}$	40	280	8	6	143
BT_{YFS}	89	166	14	1	53
$SKAT_{YFS}$	20	223	6	7	137
$SKATO_{YFS}$	47	321	7	0	140
S-PrediXcan$_{CMC}$	42	43	7	0	38
S-PrediXcan$_{METSIM}$	41	44	8	1	30
S-PrediXcan$_{NTR}$	48	70	14	6	59
S-PrediXcan$_{YFS}$	83	128	29	2	72
S-TWAS$_{CMC}$	33	45	6	0	43
S-TWAS$_{METSIM}$	36	29	5	1	20
S-TWAS$_{NTR}$	37	54	13	6	46
S-TWAS$_{YFS}$	64	105	29	2	58
OT	133	522	17	6	166
Overall	271	559	45	16	167

Notes: SCZ1 indicates the number of genes identified by each method for SCZ1 data; SCZ2 indicates the number of genes identified by each method for SCZ2 data; $SCZ_{overall}$ indicates the number of overlapping genes identified by both SCZ1 and SCZ2 data sets; $GWAS_{SCZ1}$ and $GWAS_{SCZ2}$ indicate the numbers of genome-wide significant genes that are reported in the GWAS catalog and are also identified by each method for SCZ1 and SCZ2, respectively.

Table 3.4. Genes identified only by Overall based on the two SCZ data sets that are reported in the GWAS catalog.

Gene	**Data**	**Overall**	**Reference**
RAI1	SCZ1	2.63E-31	[148]
SLC7A6	SCZ1	2.17E-15	[149,150]
AP001931.2	SCZ1	1.27E-13	[147-155]
MARK2	SCZ1	2.64E-07	[151]
GULOP	SCZ1	1.24E-07	[148-151,156]
ZBED4	SCZ1	9.02E-07	[151]
RAB11FIP5	SCZ2	1.05E-06	[151,156]
AL669918.1	SCZ2	2.03E-06	[151]
YPEL1	SCZ2	2.80E-06	[151]
LINC00606	SCZ2	2.57E-06	[151]
ERLIN1	SCZ2	2.34E-06	[151]
AC024597.1	SCZ2	2.56E-06	[152]

3.4.2 Application to the lipids GWAS summary data

We consider two lipids GWAS summary data sets, HDL1 and HDL2, which can be downloaded at the Center for Statistical Genetics (CSG) at the University of Michigan. HDL1 is a meta-analysis of HDL GWAS data set with about 100,000 samples downloaded at the website (http://csg.sph.umich.edu/willer/public/lipids2010/)[157]. HDL2 is the follow-up data with about 189,000 samples for partial validation downloaded at the Global Lipids Genetics Consortium (http://csg.sph.umich.edu/willer/public/lipids2013/)[158]. We perform the same analysis as we did in the previous section for the two SCZ GWAS summary data sets. There are 17,389 genes in HDL1 and 16,917 genes in HDL2. **Table 3.5** (HDL1 and HDL2) shows the number of genes identified by each method for the two lipids data sets, respectively. As we can see from **Table 3.5**, among the three traditional gene-based association tests with and without eQTL - derived weights, $SKATO_0$ and BT_0 identify the most number of genes in HDL1 and HDL2, respectively; Among the four S-PrediXcan tests, S-PrediXcan$_{YFS}$ and S-PrediXcan$_{CMC}$ identify the most number of genes in HDL1 and HDL2, respectively; for the four S-TWAS tests, S-TWAS$_{YFS}$ and S-TWAS$_{CMC}$ identify the most number of genes in HDL1 and HDL2, respectively. For the HDL1 data set, Overall identifies the greatest number of genes (249), followed by OT that identifies 233 genes; for the HDL2 data set, BT_0 identifies the greatest number of genes (836), followed by Overall and OT, where Overall identifies 765 genes and OT identifies 688 genes. In **Figure 3.4**, we compare genes identified by $SKATO_0$, S-PrediXcan$_{YFS}$, and S-TWAS$_{YFS}$, along with Overall and OT for the HDL1 data set and genes identified by BT_0, S-PrediXcan$_{CMC}$, S-TWAS$_{CMC}$, Overall, and OT for the HDL2 data set. Again, we observe that Overall identifies the greatest number of genes for the HDL1 data set and the second most for the HDL2 data set; all genes identified by OT are also identified by Overall; 82 and 24 genes are identified only by Overall and OT for the HDL1 and HDL2 data sets, respectively; there are 13 and 6 genes only identified by Overall for the HDL1 and HDL2 data sets, respectively. We search the GWAS catalog (https://www.ebi.ac.uk/gwas/). **Table 3.6** shows that five out of 13 genes identified only by Overall based on HDL1 data have been reported, and one out of 6 genes has been reported on HDL2 data in the GWAS catalog. We also use these two HDL GWAS data sets for partial validation by looking for the number of overlapping genes identified by both of the data sets (**Table 3.5**, HDL$_{overlap}$). There are 177 overlapping genes identified by Overall for both SCZ1 and SCZ2 data sets and 167 overlapping genes identified by OT for both SCZ1 and SCZ2 data sets.

Same as the analyses for the SCZ GWAS summary data sets, we search for genome-wide significant SNPs ($p < 5\times10^{-8}$) from the two lipids GWAS summary statistics. There are 1,911 genome-wide significant genes for HDL1 and 2,682 genome-wide significant genes for HDL2. **Table 3.5** (GWAS$_{HDL1}$ and GWAS$_{HDL2}$) summarizes the numbers of genome-wide significant genes that are identified by each method for the two lipids data sets. Among the 1,911 genome-wide significant genes for the HDL1 data set, Overall identifies the greatest number of genes (122), followed by OT (120), then $SKAT_0$ (104); S-TWAS$_{YFS}$ only identifies 29 genes and S-PrediXcan$_{YFS}$ identifies 31 genes. Meanwhile, among 2,682 genome-wide significant genes for HDL2, Overall identifies the greatest number of genes (192); OT and $SKATO_0$ identify 190 genes; S-TWAS$_{METSIM}$ and S-PrediXcan$_{METSIM}$ identify 112 and 118 genes. respectively.

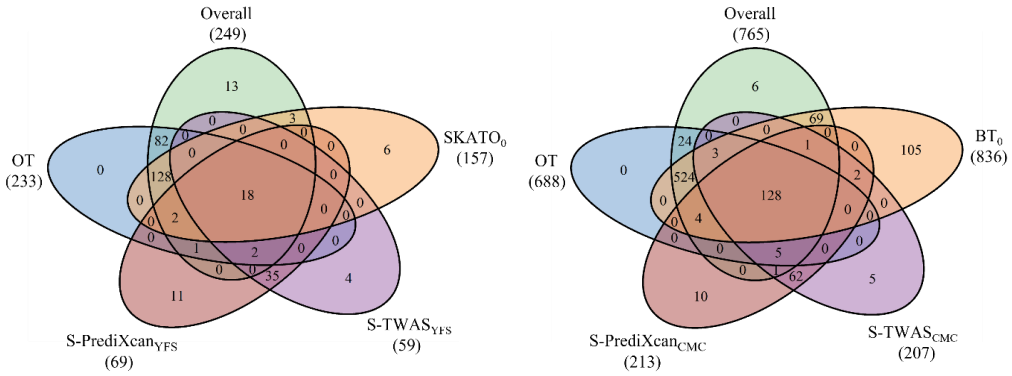


Figure 3.4. Venn diagram of the number of genes identified by Overall, OT, $SKATO_0$, S-$PrediXcan_{YFS}$, and S-$TWAS_{YFS}$ for HDL1 data (left) and HDL2 data (right). The number below each of the methods indicates the total number of significant genes identified by the corresponding method.

Table 3.5. The number of genes identified by each method for the two lipids data sets. The subscript denotes BT, SKAT, and SKATO using eQTL - derived weights; CMC, METSIM, NTR, and YFS indicate the resources to obtain the eQTL - derived weights. 0 indicates the methods without using any weights.

	HDL1	**HDL2**	**$HDL_{overlap}$**	**$GWAS_{HDL1}$**	**$GWAS_{HDL2}$**
BT_0	95	836	78	50	185
$SKAT_0$	116	174	114	99	157
$SKATO_0$	157	762	138	104	190
BT_{CMC}	79	130	41	46	107
$SKAT_{CMC}$	105	159	99	95	146
$SKATO_{CMC}$	130	177	103	96	150
BT_{METSIM}	83	160	59	58	111
$SKAT_{METSIM}$	120	259	118	102	149
$SKATO_{METSIM}$	131	199	118	98	152
BT_{NTR}	78	136	50	49	111
$SKAT_{NTR}$	105	156	100	90	148
$SKATO_{NTR}$	131	183	111	95	154
BT_{YFS}	88	154	50	53	113
$SKAT_{YFS}$	106	148	102	94	137
$SKATO_{YFS}$	142	185	112	99	144
S-$PrediXcan_{CMC}$	43	213	18	29	114
S-$PrediXcan_{METSIM}$	45	201	23	30	118
S-$PrediXcan_{NTR}$	33	187	14	19	108
S-$PrediXcan_{YFS}$	69	195	25	31	117
S-$TWAS_{CMC}$	40	207	17	23	109
S-$TWAS_{METSIM}$	37	202	16	15	112
S-$TWAS_{NTR}$	25	176	10	11	97
S-$TWAS_{YFS}$	59	183	24	29	115
OT	233	688	167	120	190
Overall	249	765	177	122	192

Notes: HDL1 indicates the number of genes identified by each method for HDL1 data; HDL2 indicates the number of genes identified by each method for HDL2 data; $SCZ_{overall}$ indicates the number of overlapping genes identified by both SCZ1 and SCZ2 data sets; $GWAS_{HDL1}$ and $GWAS_{HDL2}$ indicate the numbers of genome-wide significant genes that are reported in the GWAS catalog and are also identified by each method for HDL1 and HDL2, respectively.

Table 3.6. Genes identified only by Overall based on the two lipids data sets that are reported in the GWAS catalog.

Gene	Data	Overall	Reference
AP002954.1	HDL1	2.27E-11	[159]
EDC4	HDL1	1.65E-11	[160-162]
PACSIN1	HDL1	2.24E-06	[163]
AFF1	HDL1	2.10E-06	[164-168]
AC106779.1	HDL1	2.85E-06	[169]
NHLRC2	HDL2	1.98E-06	[166,168,170-173]

3.5 Discussions

In this paper, we develop a powerful and computationally efficient method, Overall, for gene-based association studies using GWAS summary data. Overall aggregates information from three traditional types of gene-based association tests (BT, SKAT, SKATO) and also incorporates eQTL data. Both our simulation studies and real data analysis confirm that our proposed method can control type I error rates correctly and has very good performance compared with other comparison methods. In real data analysis, Overall identify more significant genes than other methods, and there are some genes reported by GWAS catalog which are only identified by Overall.

There are some advantages of our proposed method. First, Overall adaptively aggregates information from multiple gene-based association tests. Most combination tests (i.e., Fisher's combination test[174]) assume that the p-values should be calculated from independent tests. To combine information from highly correlated gene-based association tests, Overall utilizes the extended Simes procedure[119,135]. It is shown that this procedure to combine multiple tests is stable and effective regardless of whether the tests are highly correlated[137,175]. Second, Overall is more powerful than the traditional gene-based association tests, some popular transcriptome association tests (i.e., S-PrediXcan[142] and S-TWAS[116]), and other eQTL weighted combination tests (i.e., ominous test[128]). By aggregating information from different tests and incorporating multiple eQTL - derived weights, Overall can achieve a higher statistical power under a variety of situation settings. Meanwhile, our simulation studies and real data analyses show that the extended Simes procedure is more powerful than the Cauchy combination method, especially if the proportion of causal variants in a gene is small. Third, the p-values of Overall can be analytically computed without using permutations, therefore, Overall is computationally efficient. Finally, using the regularization procedure to correct the estimated LD can reduce the potential statistical noise in the LD estimation if LD is estimated using a reference panel with small sample size. In addition, Overall can be easily applied to genetic association studies with either individual-level data or GWAS summary statistics.

In this paper, we combine three types of traditional gene-based association tests (BT, SKAT, SKATO). However, the combination procedure used in the paper is very general. Other more powerful gene-based association tests can also be combined using the same approach, such as some state-of-the-art methods (i.e., S-TWAS[116], E-MAGMA[176], and SMR[177]).

In this current study, we utilize the weights derived from four single tissue gene expression studies (CMC, METSIM, NTR, YFS). Although the extended Simes procedure in Overall allows us to employ more eQTL – derived weights from a number of studies (i.e., GTEx gene expression version 8[178] et al.), there is a possibility that the noise can be increased with the increment in the number of unrelated studies. Therefore, the power of the combination tests (i.e, Overall and OT) might be attenuated. Thus, to obtain the most robust identification of phenotypic associated genes in a real data analysis with the Overall method, we suggest incorporating eQTL datasets from the most relevant tissues to the phenotype. The last but the most important thing is that population stratification can be confounded association results[50,179]. Systematic minor allele frequency difference between transcriptomic studies of different cohorts and no matching between the estimated LD structure of Genomes Project with that in the study may increase the chances of false positive findings. Therefore, we need to eliminate false positive findings possibly caused by population stratification[180,181]. When applying the Overall method, the population of GWAS summary dataset, external reference panel (i.e., 1000 Genomes Project) used to estimate LD structure, and eQTL – derived weights should be consistent.

In this study, the computational time of the proposed method is acceptable even if the estimated correlation matrix of multiple tests is obtained by the replication procedure. Meanwhile, the estimation procrdure is independent of gene-based association tests, therefore we only need to perform this procedure once for each GWAS summary dataset. For example, there are a total of 29,008 gene in the 1000 Genomes Project and we use 1,000 replicates to estimate the correlation matrix of multiple tests for each gene. We perform this using the high-performance computing (HPC) cluster (Intel Xeon E5 – 2670 2.6 GHz, 16 GB RAM). The computational time for all genes is about 36 hr CPU time with 500 nodes. Then, the p-value of the proposed method can be computed analytically which is independently performed in each GWAS summary dataset. The computational time for each GWAS dataset is about 1 hr CPU time with 10 nodes.

4 Chapter 4

TGPred: Efficient methods for predicting target genes of a transcription factor by integrating statistics, machine learning, and optimization

Abstract

Six statistical selection methods were developed based on the penalized regression models with two loss functions (mean squared error (MSE) and Huber function (Huber)), and three penalty functions (Lasso, elastic net (ENET) and network-based penalty (Net)), for inferring target genes of a transcription factor (TF) of interest. We also ameliorated an accelerated proximal gradient descent (APGD) algorithm to optimize parameter selection processes of the six methods, resulting in a much more efficient APGD algorithm than the commonly used convex optimization solver (CVX). As the synthetic data generated from the general setting was used to test four non-Net methods, MSE-ENET penalty performed better while Huber-Lasso performed worse than other methods. As the synthetic data generated from the network setting was used to test all six methods, MSE-Net and Huber-Net outperformed the non-Net methods. The non-Net methods were also tested with SND1 and GL3 overexpression real transcriptomic data sets. Huber-ENET and MSE-ENET outperformed Huber-Lasso and MSE-Lasso in overall. The methods we developed will fill the gap of lacking the appropriate methods for predicting target genes of a TF, and are instrumental for validating experimental results yielding from ChIP-seq and DAP-seq, and conversely, selection and annotation of TFs based on their target genes.

Keywords: transcription factors, target gene prediction, selection probability, statistical selection, and convex optimization

4.1 Introduction

Construction and delineation of transcriptional regulatory networks are essential for systematically understanding how various biological processes and complex traits are regulated at system level and how plants grow and develop in response to environmental cues. Although biological experiments can be performed to obtain gene regulatory relationships, they are labor-intensive and time-consuming, and are only applicable to acquire a small number of true regulatory relationships due to a tremendous amount of work. In the last two decades, the advent of high throughput technologies including microarray, RNA-Seq, and ChIP-seq as well as DAP-seq, made it easier to generate a terabyte transcriptome data for network inference. As the high-throughput data in public repositories increase exponentially, various computational algorithms and tools utilizing high-throughput transcriptome data and ChIP/DAP-seq provide an alternate approach to infer gene regulatory relationships and acquire gene regulatory networks. However, the acquisition of transcriptional gene regulatory network with high accuracy is pivotal for

such an approach. To develop highly accurate methods, exploration of machine learning, statistics and optimization combined approaches is promising and opens a new avenue for doing this more efficiently.

In the earlier stage, for example, one to two decades ago, high-throughput transcriptome data were primarily generated from single cell organisms like bacteria and yeast, or the cell lines of eukaryotic organisms, which allowed to generate time-course microarray data with small time intervals. These types of data encouraged the development of many dynamic methods that incorporated the temporal variable into the models to accurately predict gene regulatory relationships, such as differential equations[182], finite state[183], dynamic Bayesian[184], Boolean network[185], and stochastic networks[186] and ordinary differential equations (ODE)[187]. For these methods, the time-course data with very small time intervals are critically important to the accuracy of inferred networks and the regulatory relationships therein contained. Since it is very time-consuming to harvest specific cell types or tissues from the multi-cellular organisms, more and more high-throughput transcriptome data were generated from various tissues of multicellular organisms like plants and mammals in a loosely timed series or entirely no points. Static data are thus characterized by very large time intervals (e.g. days or weeks) or non-time-points at all. To analyze this kind of data, the static methods, which do not involve temporal variable, were developed, such as ParCorA[188], maximum relevance/minimum redundancy Network (MRNET)[189], mutual information based relevance networks[190], Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE)[191], Context Likelihood of Relatedness (CLR)[192], C3NET[193], Mutual Information 3 (MI3)[194], and probabilistic-based Bayesian network[195], random forests[196].

Recently, more methods have been developed for constructing local gene regulatory networks especially the multilayered gene regulatory network, such as top-down GGM[197,198], bottom-up GGM algorithm[199], and BWERF[200], and gene regulatory network controlling a pathway or a biological process, for instance, TGMI[201] and HB-PLS[202]. In addition, the methods for constructing multiple joint gene regulatory networks using data from multiple sources, for example, JGL[203] and JRmGRN[204], have been developed. However, the above-mentioned methods are not specifically tailored to the needs of inferring the target genes of a transcription factor (TF). In reality, we desperately need the methods for inferring the targets genes of each TF for facilitating construction of a complete network and validating regulatory relationships or the networks inferred based on *in-silico* analyses and biological experiments. For example, we need the methods of inferring the targets genes of a TF of interest, which can be employed to validate experimental results of ChIP-seq and DAP-seq. Conversely, such methods allow us to infer a TF's functions based on the functions of its target genes. After multiple TFs' targets genes are inferred, we can screen TFs for specific purposes based on their target gene functions.

In this study, we developed six statistical selection methods to infer the potential TGs for a given TF, which combined two loss functions and three penalty functions. The loss functions, mean squared error (MSE) and Huber function (Huber), were used to measure the errors between the predicted values and the observed values. Huber can avoid the sensitivity of heavy-tailed errors or outliers than MSE. The penalty functions, Lasso, elastic net (ENET) and network-based penalty (Net), contain the l_1 norm of the estimated

effect sizes which can control the sparsity of the selected TGs. Meanwhile, Net penalty can incorporate prior biological genetic network information into the prediction [205]. We also modified and implemented an accelerated proximal gradient descent (APGD) algorithm for the parameter optimization in all six methods. Our simulations showed that the APGD was much more efficient than a commonly used method called convex optimization solver (CVX). To obtain a stable selection result, we applied the stability selection method, namely, half-sample approach, which does not need to choose the optimal tuning parameters in selection methods. We tested all the methods with simulated data, and four non-Net methods with the real transcriptomic data of all genomic genes, and two Net methods with the real transcriptome data of all metabolic pathways, especially lignin pathway genes. Our study showed that the four non-Net methods were useful for identifying the target genes of a TF of interest in genome-wide analysis, which implies that the methods could be used to validate target genes of a TF resulting from TF ChIP-seq or DAP-seq experiments, while the two Net-based methods can identify TGs involved or associated with a pathway or a biological process, and TFs that regulate them. When multiple TFs are analyzed, the results can be used for TF selection and screening based on the distinct functions of their target genes.

4.2 Materials and Methods

4.2.1 Materials

Simulated gene expression data

The simulated data were generated in two settings: (1) general setting; (2) network setting. In the general setting, p TGs were independent with each other and the first 50 TGs were regulated by a given TF (Details in **Text D.1**). In the network setting, we simulated p TGs with two biological network structures, the hierarchical network and Barabasi-Albert network. For the hierarchical network, there were 5 disjointed subnetworks and each of them consisted of 100 TGs. The subnetwork was constructed as the same as Kim et al.[205] (**Figure D.1**). For the Barabasi-Albert (BA) network, there were 50 subnetworks and each of them was a BA-based network comprising of 10 TGs[206]. There were 45 TGs and 40 TGs that were regulated by a given TF for the hierarchical network and Barabasi-Albert network, respectively (Details in **Text D.2**).

Populus trichocarpa SND1 transcriptomic data and analysis

The poplar data used for simulation were from our previous studies[197]. The data can be retrieved from Gene Expression Omnibus (GEO) with accession number GSE49911. Briefly, the data were generated and then analyzed as following: Poplar protoplasts isolated from stem developing xylem were transfected with plasmid vector harboring poplar SND1 gene under the control of 35S promoter, and then harvested for RNA-seq at 7, 12 and 25 hours. Three samples of SND1-driven by 35S at each time point were harvested while three control samples (control vector without SND1) at each time point were harvested. The raw count data were used for identification of differentially expressed genes (DEGs) for each time point using the edgeR package[207], and for normalization with trimmed mean of M-values (TMM) contained in the edgeR package. Normalized data were used for real data simulation to validate the methods we developed in this study.

Maize gl3 transcriptomic data and analysis

Two transcriptional-activator like effectors (dTALes) that target two non-overlapping 16-bp regions of the *gl3* promoter for overexpression were constructed. The two regions are located 5 bp and 48 bp upstream of the transcription start site. 14 day-old seedlings were used to test for *gl3* dTALes-mediated induction of *gl3*, and bacterial strains carrying either dT1 or dT2 activated *gl3* expression by 24 hours after the bacterial inoculation. Three samples and three controls, upon being infected with Xv1601 bacteria carrying dTALes, were harvested in a time-series with four time points: 6, 12, 24, and 48 hours. Sequencing data were trimmed by Trimmomatic (version 0.38)[208]. Trimmed reads were aligned to the maize B73 reference genome (B73Ref4) using STAR (2.7.3a)[209]. The data were aligned to maize genome B73 from which FPKM values were generated with Cufflink package[210], and DEGs were identified with Cuffdiff package[211]. FPKM data were used for simulation with *gl3* as TF and DEGs and all genomics genes as candidate TGs. The dTALe RNA-Seq data are available at NCBI SRA under the project of PRJNA692729.

Maize B73 transcriptomic data for validation of Net-based methods

In total, the expression levels of 739 RNA-seq data of B73 were downloaded from NCBI Sequence Read Archive (SRA) repository. The accession numbers are shown in Table S1. Raw read counts generated per gene were calculated by STAR and then normalized with Cufflink[212]. 2,539 unique pathway genes were extracted from the Plant Metabolic Network (PMN)[213] and 23 lignin pathway genes as well as 23 transcription factors (TFs) that are known to regulate lignin pathway[214-218] were used for validating the Net-based methods, Huber-Net and MSE-Net.

4.2.2 Statistical selection methods

Consider that the expression levels of a TF $\boldsymbol{y}$ and the expression levels of the TGs $\boldsymbol{x}$ in the whole-genome form a linear relationship:

$$y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \;\; i = 1, \cdots, n,$$

where n is the number of samples, $\boldsymbol{x}_i = \left(x_{i1}, \cdots, x_{ip}\right)^T$ is the expression levels of p TGs in sample i, and y_i is the expression level of the TF gene in sample i. β_0 is the intercept and $\boldsymbol{\beta} = \left(\beta_1, \cdots, \beta_p\right)^T$ are the regulated regression coefficients. The TF gene regulates TG *j* if $\beta_j \neq 0$ $(j = 1, \cdots, p)$; the TG *j* and TG k are co-regulated by TF gene if both $\beta_j \neq 0$ and $\beta_k \neq 0$. ε_i is independent and identically distributed random errors with mean 0 and variance σ^2.

Based on the above statistical model, we developed six statistical selection methods to infer the potential TGs for a given TF based on the penalized regression model. The general objective function of the penalized regression model was defined as

$$f(\boldsymbol{\beta}; \lambda, \alpha) = L(\boldsymbol{\beta}; y, \boldsymbol{x}) + P(\boldsymbol{\beta}; \lambda, \alpha),$$

where $L(\boldsymbol{\beta}; y_i, \boldsymbol{x}_i)$ is the loss function according to the observed expression levels of TGs and TF and $P(\boldsymbol{\beta}; \lambda, \alpha)$ is the penalty function which can control the sparsity of the selected TGs.

Loss functions

In the above general objective function of the penalized regression model, we considered the following two loss functions, MSE and Huber. The MSE loss function is defined as $L^{MSE}(\boldsymbol{\beta}; y, \boldsymbol{x}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$, which is very sensitive to outliers. Therefore, the use of Huber loss function has been proposed and is more robust to the heavy-tailed errors or outliers than MSE[219]. The Huber loss function is defined as $L^{Huber}(\boldsymbol{\beta}; y, \boldsymbol{x}) = \sum_{i=1}^{n} H_M(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})$, where $H_M(z)$ is the Huber function for an input value z, which is quadratic function for small z values but grows linearly for large values of z. In this study, the parameter M is defaulted to be one-tenth of the interquartile range (IRQ), as suggested by Deng et al.[202]. For any given positive real M (called shape parameter), the Huber function is defined as

$$H_M(z) = \begin{cases} z^2 & |z| \le M \\ 2M|z| - M^2 & |z| > M \end{cases}$$

Penalty functions

All of the three penalty functions we considered, Lasso, ENET, and Net, contained the l_1 norm of the estimated effect sizes ($\|\boldsymbol{\beta}\|_1$). The ENET penalty is the combination of the l_1 norm and squared l_2 norm, $P^{ENET}(\boldsymbol{\beta}; \lambda, \alpha) = \lambda\alpha\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda(1-\alpha)\|\boldsymbol{\beta}\|_2^2$. $\lambda > 0$ and $\alpha \in [0,1]$ are the tuning parameters, where λ controls the sparsity and α is the mixing proportion between l_1 norm and l_2 norm. The Lasso penalty is the special case of ENET ($\alpha = 1$) and $P^{Lasso}(\boldsymbol{\beta}; \lambda, \alpha) = \lambda\|\boldsymbol{\beta}\|_1$, which only contains one tuning parameter $\lambda > 0$. For the Net penalty, we assume that the genes involved in the same pathway are often co-regulated by a TF or the same regulatory mechanism, which is supported by previous studies[220-222]. The Net penalty function can utilize prior biological network knowledge such as genetic pathways[205], which is a combination of the l_1 norm and squared l_2 penalty using the genetic network structure. As introduced in Kim and Sun[205], the $P^{Net}(\boldsymbol{\beta}; \lambda, \alpha)$ is defined as

$$\begin{aligned} P^{Net}(\boldsymbol{\beta}; \lambda, \alpha) &= \lambda\alpha\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\mathbf{S}^T\mathbf{L}\mathbf{S}\boldsymbol{\beta} \\ &= \lambda\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1}{2}\lambda(1-\alpha)\sum_{j=1}^{p}\sum_{j\sim k}\left(\frac{s_j\beta_j}{\sqrt{d_j}} - \frac{s_k\beta_k}{\sqrt{d_k}}\right)^2 \end{aligned}$$

In the above formula, $\mathbf{S} = diag(s_1, \cdots, s_p)$ is a diagonal matrix whose diagonal entries are the signs of estimated regression coefficients, which can be obtained from either the ordinary regression when $p < n$, or the ridge regression when $p \ge n$. It has been shown that the matrix **S** can accommodate the problem of failure of local smoothness between linked genes[223]. For example, if two nearby TGs are negatively regulated by TF, the signs in their regression coefficients are expected to be different. **L** is a symmetric normalized Laplacian matrix, where the elements of **L,** L_{jk}, are given by

$$L_{jk} = \begin{cases} 1 & if\ j = k\ and\ d_j \neq 0 \\ -(d_j d_k)^{-\frac{1}{2}} & if\ j \neq k\ and\ j\sim k \\ 0 & otherwise \end{cases},$$

where $j\sim k$ means that the TGs j and k are linked in the genetic network and d_j is the total number of genes linked with the TG j. Note that the genetic network information **L** are considered as the functional relationships among the TGs, which can be obtained from the existing annotation. For example, we can construct an association network using the pathways or biological processes information, where the TGs are associated with each other if they are within a metabolic pathway or a biological process.

Based on the above two loss functions along with three penalty functions, we developed six statistical selection methods, named MSE-Lasso, MSE-ENET, MSE-Net, Huber-Lasso, Huber-ENET, and Huber-Net. For a given pair of λ and α, we can estimate the regression coefficients of p TGs, $\boldsymbol{\beta}$, by minimizing the objective function $f(\boldsymbol{\beta};\lambda,\alpha)$ introduced in formula (2). In other words, $\boldsymbol{\beta}=\text{argmin}_{\boldsymbol{\beta}}f(\boldsymbol{\beta};\lambda,\alpha)$. The penalty function $P(\boldsymbol{\beta};\lambda,\alpha)$ is convex[205,224], so the solution to $\boldsymbol{\beta}$ can be obtained via one of the convex optimization algorithms.

4.2.3 Algorithm to solve the penalized regression models

Since $|\beta_j|$ is convex but not differentiable at $\beta_j=0$ for $j=1,\cdots,p$, it is difficult to use the gradient descent method to find $\boldsymbol{\beta}=\text{argmin}_{\boldsymbol{\beta}}f(\boldsymbol{\beta})$. Although here we can use the general convex optimization solver CVX[225], it is too slow for real biological applications especially when there are a large number of genes involved in the analysis. Therefore, we adapted an accelerated proximal gradient descent (APGD) algorithm which is an effective algorithm when the objective function can be decomposed as a sum of a convex differentiable function and a convex non-differentiable function. In the six methods we developed, the objective function $f(\boldsymbol{\beta})$ can be decomposed as $f(\boldsymbol{\beta})=g(\boldsymbol{\beta})+h(\boldsymbol{\beta})$, where $g(\boldsymbol{\beta})$ is a convex differentiable function and $h(\boldsymbol{\beta})$ is a convex non-differentiable function. The idea behind APGD method is to make a quadratic approximation to $g(\boldsymbol{\beta})$ and leave $h(\boldsymbol{\beta})$ unchanged [226], then use the iterations to solve $\boldsymbol{\beta}=\text{argmin}_{\boldsymbol{\beta}}f(\boldsymbol{\beta})$ (Details in the **Texts D.3-D.8**).

4.2.4 Selection probability

To obtain a stable selection result, we applied the stability selection method, namely, half-sample approach, to each TG, which does not need to choose the optimal tuning parameters in selection methods. For a pair of fixed values of λ and α ($\alpha=1$ for Lasso penalty), $n/2$ samples are selected at random without replacement and then the regression coefficients are estimated based on this subset of samples. This process is repeated B times for each pair of α and λ over a grid set of α and λ. Let $\hat{\beta}_j(S_b;\alpha,\lambda)$ denote the estimated regression efficient for the bth sample ($S_b, b=1,\cdots,B$), the selection probability of TG j, SP_j, is the maximum portion of non-zero $\hat{\beta}_j(S_b;\alpha,\lambda)$ over all pairs of α and λ. In other words,

$$SP_j=\max_{\alpha,\lambda}\frac{1}{B}\sum_{b=1}^{B}I\left(\hat{\beta}_j(S_b;\alpha,\lambda)\neq 0\right)$$

where $I\left(\hat{\beta}_j(S_b;\alpha,\lambda)\neq 0\right)$ is an indicator function and $I\left(\hat{\beta}_j(S_b;\alpha,\lambda)\neq 0\right)=1$ if $\hat{\beta}_j(S_b;\alpha,\lambda)\neq 0$ for $b=1,\cdots,B$.

There are two major advantages for the use of selection probability. First, we avoid selecting the optimal tuning parameters λ and α, which is challenging in penalized regression analysis. Second, it has been shown that the results obtained from the half-sample approach and the selection probability are more stable than those obtained from the cross-validation[205,227]. The main challenge of the stability selection method is how to appropriately choose the grid sets of the two parameters λ and α. For a given α, the smallest λ such that all estimated coefficients are zeros from two loss functions, MSE and Huber, can be defined as

$$\lambda_{max}^{MSE} = \max_{j=1,\cdots,p} \left| \sum_{i=1}^{n} (y_i - \beta_0 - x_{ij}\beta_j) x_{ij} \right| / \alpha,$$

$$\lambda_{max}^{Huber} = \max_{j=1,\cdots,p} \left| \sum_{i=1}^{n} \nabla H_M(y_i) x_{ij} \right| / \alpha$$

where $\nabla H_M(y_i) = 2y_i I(|y_i| \le M) + 2Msign(y_i) I(|y_i| > M)$ is the gradient of Huber function. Therefore, the grid set of λ can be set as a $\log_{10}$-scale from $ratio * \lambda_{max}$ to λ_{max}, where the $ratio = 0.01$ as suggested by R package *glmnet*.

Six statistical selection methods based on the penalized regression models and the APGD algorithm for solving these six statistical methods had been implemented in both Python3 and R and then packed into TGPred, which have been made publicly available on GitHub as open-source software for downloading (https://github.com/xueweic/TGPred); more detailed information on how to install and run the tool was enclosed in the packages; also see **Text D.9**.

4.3 Simulation studies

Simulation studies were used to evaluate the performance of the six statistical selection methods we developed based on the penalized regression models. We considered two simulation settings, the general setting and the network setting, and we used $n = 300$ samples and $p = 500$ TGs in all simulation settings. For each simulation setting, the regulation effects for all genes based on each method were estimated by APGD, and the selection probabilities were calculated by $B = 500$ half-sample approach. Then, the true positive rates (TPRs) were used to evaluate the selection performance, which is defined as the number of the truly regulated genes among the selected top-ranked genes divided by the total number of truly regulated genes.

In the general setting, TGs were independent with each other. Therefore, we only compared the performances of Huber-Lasso, MSE- Lasso, Huber-ENET, and MSE-ENET in the general setting. **Figure 4.1** showed the TPRs of these for methods in the general setting based on the number of selected top-ranked genes. As it is known, the bigger pre-set regulation effects may result in the higher TPRs of all methods, since all methods can select the genes with larger true regulation effects. On the contrary, the lower pre-set regulation effects may result in the lower TPRs of all methods. In both cases, we cannot differentiate the performances of different methods. Therefore, we pre-set the regulation effects $\beta = 0.2$ or 0.3, and 50 TGs were regulated by a given TF in this simulation setting. For $\beta = 0.3$, all four methods achieved over 80% TPRs when we selected 50 top-ranked genes, while all of them performed equivalently well when we selected 40 top-ranked

genes or less. After selected 85 top-ranked genes, all methods achieved over 95% TPRs and MSE-ENET performed better than the other three methods. Compared with Huber loss function, MSE loss function had higher TPRs no matter what penalty functions were used. The Area under the Receiver Operating Characteristic curve (AuROC) measured the performance across all possible thresholds of selection probabilities. Note that the larger the AuROC, the better the performance of the method. All of four methods obtained an AuROC that exceeded 0.9. As shown in **Figure D.2**, AuROC (MSE-ENET) = 0.97, AuROC (MSE-Lasso) = 0.95, AuROC (Huber-ENET) = 0.95, and AuROC (Huber-Lasso) = 0.91. Similar to $\beta = 0.3$, MSE-ENET performed best and all methods achieved over 70% TPRs when we selected 50 top-ranked genes along with over 0.8 AuROC.

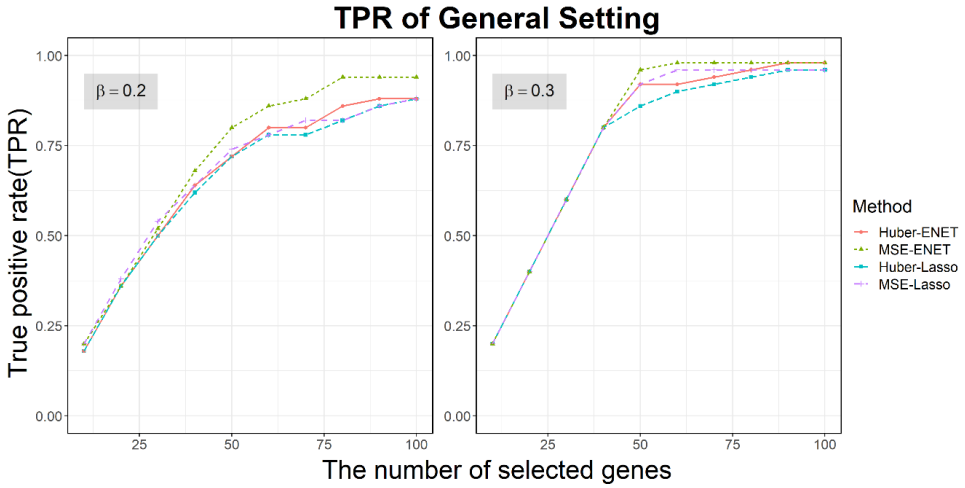


Figure 4.1. The TPRs of different methods in general setting. The selection probabilities were calculated using half-sample approach method with $B = 500$ times of resampling.

For the network setting, we considered two network structures, the hierarchical network (**Figure D.1**) and the Barabasi-Albert network (not shown). **Figure 4.2** showed how TPRs varied with the different numbers of the top-ranked genes for different methods in the hierarchical network and the Barabasi-Albert network. For the hierarchical network where 45 TGs (out of 500 genes) were truly regulated by a given TF, we pre-set the regulation effects $\beta = 0.3$ or 0.4. Since the Net penalty function incorporated the network structure, TPRs of Huber-Net and MSE-Net were higher than the other four methods. For the Barabasi-Albert network where 40 true TGs (out of 500 genes) were regulated by a given TF, we pre-set the regulation effects $\beta = 0.1$ or 0.2. Huber-Net and MSE-Net had the highest TPRs in all settings as expected, indicating that Huber-Net and MSE-Net have the same performances and outperform the other four non-Net methods in both network structures. We then plotted the ROC curves for all methods in two network settings. For the hierarchical network setting, the AuROCs of both Huber-Net and MSE-NET were 0.73 for $\beta = 0.3$, and were 0.78 for $\beta = 0.4$, which were higher than the AuROCs of the other four methods (**Figure D.3**), indicating that Huber-Net and MSE-Net can incorporate the functionally associated genes and increase the probability of these genes to be selected as the TGs for a given TF. Meanwhile, for the Barabasi-Albert network, the AuROCs of both Huber-Net and MSE-Net were 0.9 for $\beta = 0.1$, and were 0.95 for $\beta = 0.2$, which were

also higher than the AuROCs of the other four methods (**Figure D.3**). Based on both TPR and AuROC, we conclude that Huber-Net and MSE-Net performed equivalently well and out-performed all other four non-Net methods. Compared to the general setting, it is obvious that the four non-Net methods performed less differentially in the two network settings, as shown in **Figures 4.2** and **Figure D.3**.

We also compared the computation time and the regression coefficients estimated by APGD and CVX, a commonly used package for convex optimization, for several pairs of tuning parameters λ and α. The comparison results were shown in **Figures D.4-D.9**, which were also summarized into **Text D.10** for the detailed analyses. For brief, our simulation results showed that APGD was not only capable of obtaining the similar estimated regulation effects of all TGs for a given TF, but also shortened the computation time to 1/10 of that by using CVX, which enables us to predict true TGs of a TF out of a large number of candidate TGs (e.g. more than 30,000 as demonstrated in **Figure D.4**).

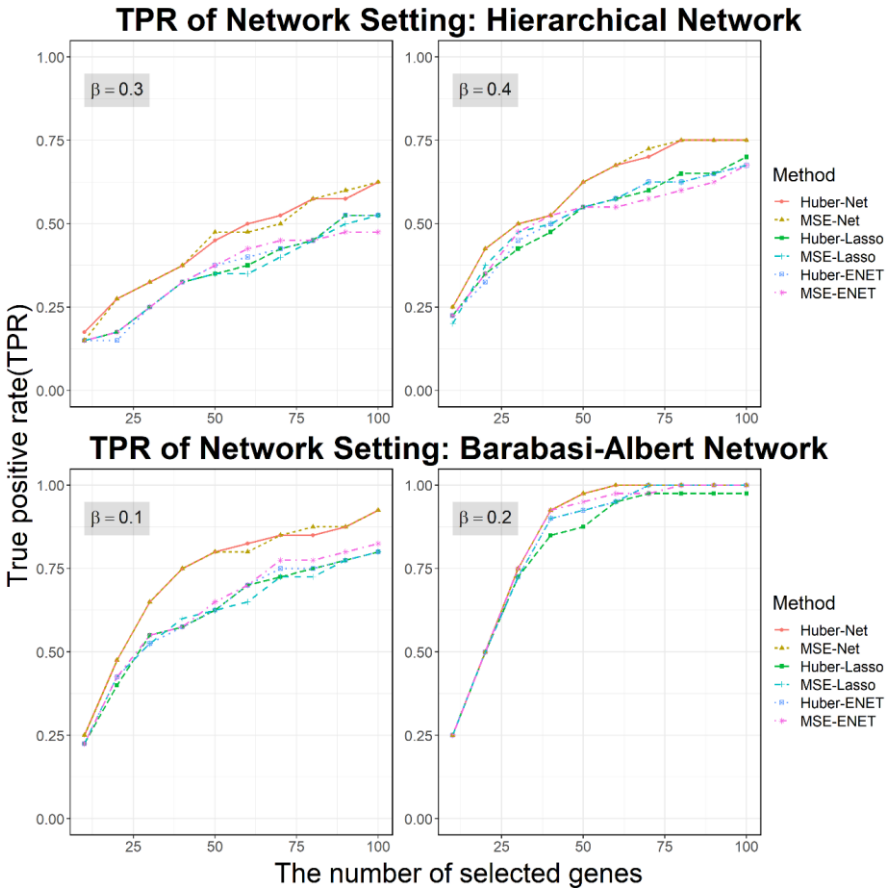


Figure 4.2. The TPRs of different methods in network setting. The selection probabilities were calculated using half-sample approach method with $B = 500$ times of resampling.

4.4 Real data analysis

4.4.1 Validating Non-Net methods with SND1 transcriptomic data

DEG analysis yielded 178 genes that had significant p-values (< 0.05) as shown in our early publication[197]. The Top-down GGM algorithm identified 94 genes that were tightly responsive to SND1, from which we identified 84 genes that were interfered by SND1 directly and hereafter, referred to as putative direct TGs of SND1. Of these 84 direct TGs, we randomly drew 16 genes for experimental validation with ChIP-PCR, all 16 genes tested are proven to be the direct TGs of SND1, which 15 of 16 genes chosen from indirectly genes ($178 - 84 = 94$) are proven to be true indirect TGs, indicating the high accuracy (93%) of the Top-down GGM algorithm. Using the same data set we then simulated SND1 and all 33,691 genomic genes, and attempted to identify the direct TGs based on the selection probabilities yielded from each method. When the all genes being ranked by selection probabilities, Huber-ENET, MSE-ENET, HuberLasso and MSE-Lasso identified 58, 53, 42 and 43 responsive genes, and 53, 49, 38 and 39 TGs, among the top 178 genes, respectively. We plotted AuROC and obtained ROC and AuROC of the four methods (**Figure 4.3**). The ROC curves showed that the Huber-ENET and MSE-ENET ranked more positive TGs to the very top of lists as compared to Huber-Lasso and MSE-Lasso, indicating that ENET penalty outperformed Lasso. Interestingly, the Huber-ENET and MSE-ENET performed very well when they were used to identify TGs of SND1 from all genomic genes all genomic genes (33, 691 genes) (**Figure 4.3** left), as evidenced by the AuROCs ≥ 0.75.

4.4.2 Validating Non-Net methods with *gl3* transcriptomic data

We employed the Transcriptional-Activator Like effectors (TALes) to activate glossy3 (*gl3*), a glossy master regulator. Two dTALes, referred to as dT1 and dT2, were constructed to target two non-overlapping 16-bp regions in the *gl3*'s promoter. The two regions targeted by dT1 and dT2 are 4 and 48 bps upstream of the transcription start site of *gl3*. Analysis of RNA-seq data yielded at 24 h revealed 144 genes (93 upregulated and 51 downregulated genes), that were activated by both dT1 and dT2[228]. From these 144 genes, we identified 93 tightly responsive genes to *gl3* and 78 TGs of gl3using Top-Down GGM Algorithm with a cut-off corrected p-values < 0.05. The 78 genes contain 6 of 9 known glossy genes in the literature, supporting that the 78 genes are true positive TGs. When we implemented the four non-Net methods we developed to *gl3* and all 30,263 genomic genes, and attempted to identify the responsive genes and TGs of *gl3* based on the selective probabilities. When the top 144 genes were ranked by selective probabilities, Huber-ENET, MSE-ENET, HuberLasso and MSE-Lasso identified 78, 81, 91, and 93 responsive genes and 57, 49, 68 and 70 TGs, respectively, among the top 144 genes. We plotted AuROC and obtained ROC and AuROC of the four methods (**Figure 4.3**). The ROC curves showed that the Huber-ENET and MSE-ENET ranked slightly more positive TGs to the top of list as compared to Huber-Lasso and MSE-Lasso, indicating ENET penalty outperformed Lasso. Intriguingly, all four non-Net methods performed very well when they were used to identify TGs of gl3 from all genomic genes (30, 263 genes) (**Figure 4.3**, right), as evidenced by the AuROCs ≥ 0.91.

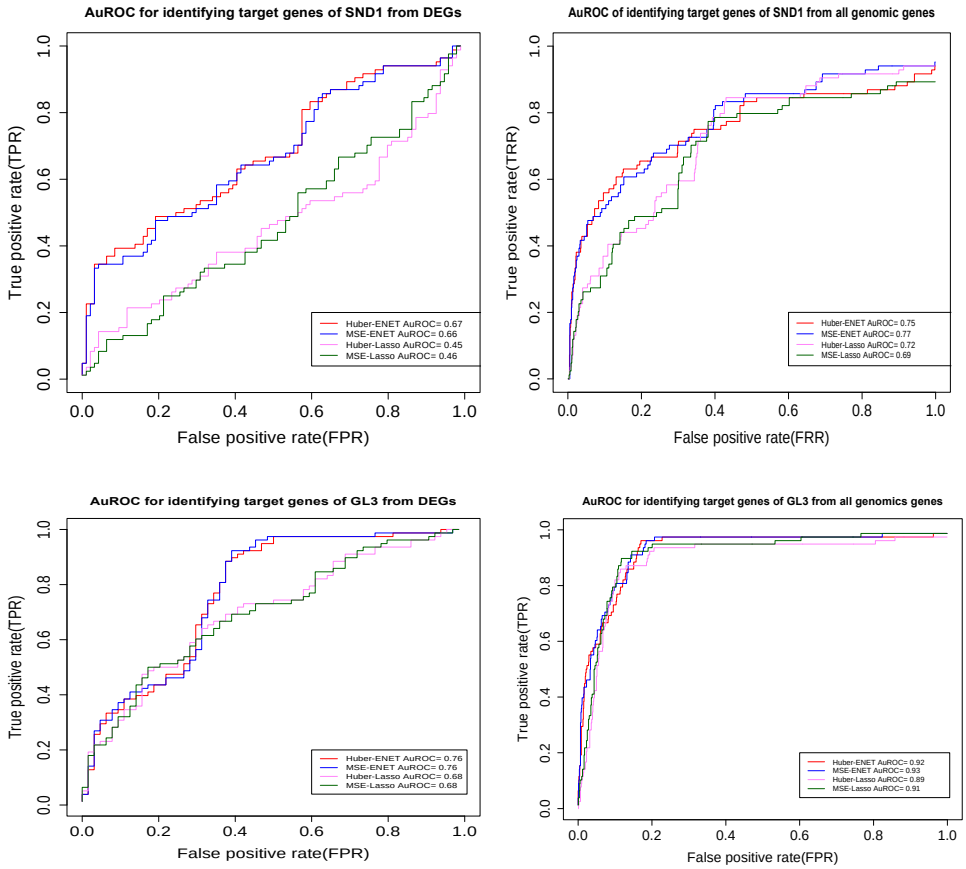


Figure 4.3. The performance of four non-Net methods in TGPred package. A. ROC generated with the data set of 178 differentially expressed genes (DEGs) of SND1 from *Populus trichocarpa*. B. ROC generated with the data set of all genes (33,691) in the RNA-seq data from *Populus trichocarpa*. C. ROC generated with the data set of 144 DEGs of *gl3* from *Zea mays*. B. ROC generated with the data set of all genes (30,263) in the RNA-seq data from *Zea mays*. DEG, differentially expressed genes. AuROC, area under the receiver-operating characteristic curve.

4.4.3 Validating Net-based methods with lignin pathway in Maize

Maize expression data has been used for predicting the regulatory relationships between transcription factor (TFs) and pathway genes (PWGs). A total of 2,539 PWGs belonging to at least one pathway were obtained after the genes that have 90% expression values are 0 in the 739 samples were removed. These 2,539 PWGs belong to 446 pathways. To evaluate the performance of our proposed six methods and APGD algorithm in real data analysis, we applied each method to each of 23 TFs versus 2,539 PWGs. The Laplacian matrix **L** of 2,539 PWGs was constructed based on 446 pathways, that is, two PWGs were associated together if they belong to at least one of 446 pathways. Since these 23 TFs are the known TFs that regulate lignin pathway in multiple plant species[229]. We specifically examined 21 genes in maize which were curated by Plant Metabolic Pathway[213] as lignin pathway genes.

We applied our proposed six methods to the 739-sample data sets of 2,539 PWGs, and 23 TFs to calculate the selection probability of 2,539 PWGs for each TF. For three penalized regression methods, HuberNet, HuberENET, and ENET, nine α values ($\alpha = 0.1, 0.2, \cdots, 0.9$) and 10 different λ values in a calculated range from the loss function ("Lambda_gird" function from our developed package "TGPred") were used, respectively. For Huber-Lasso, 100 λ values in a calculated range from the loss function with $\alpha = 1$ were used. Furthermore, the parameter *B* representing the number of subsets of samples drawn with the half-sample resampling method were used to calculate the selection probabilities of 2,539 PWGs for each TF. Then, we specifically checked 21 lignin pathway genes to verify the reliability of our methods. There are two criterions that are regarded as existing regulator relations. We chose the PWGs with the selection probabilities greater than 0.90. The PWGs were captured by the six methods were shown in **Figure 4.4A** and **Figure 4.4B**. The results yielded from three methods with Huber loss function (**Figure 4.4A**) and the results yielded from three methods with MSE loss function (**Figure 4.4B**) were placed side-by-side for comparisons. It is obvious that the values of selection probabilities calculated by three methods with Huber loss functions were larger than those of three methods with MSE loss function, as indicated by the color depths (**Figure 4.4**).

Currently, there are no methods that have been developed and tailored specifically towards identifying the TGs of a given TF, especially for a genome-wide analysis. As a result, we could not find a similar or closely related method that can be used as a comparison to illustrate the efficacy of the six methods. We finally used a widely used network construction method, ARANCE[191], as a comparison. The results yielded by ARANCE are shown in **Figure 4.4C** when the same inputs (2,539 PWGs) as six methods were fed. Nevertheless, only a few regulatory relationships were captured by the ARANCE. When we used only the 21 lignin pathway genes as the input for ARANCE, more regulatory relationships were captured (**Figure 4.4D**), which were still much less than those identified by the six methods. Compared to ARANCE, the six methods identified many additional and unique relationships.

Huber-Net identified the unique pathway genes that were not identified by Huber-ENET and Huber-Lasso. For example, *CAD1* regulated by *MYB20*, *4CL2-1* by *VND1*, *HCT1.1* by *MYB59*. However, based on the lignin pathway genes alone, the differences in target identification by the six methods were not largely different. To examine this with all pathways, we show the common and unique TGs of the same TFs of the three methods that use the same loss function (Huber or MSE) for 23 TFs versus all 2,539 pathway genes. As shown in the Venn diagram (**Figure D.10**), Huber-Net and MSE-Net identified up to 10 and 27 unique genes for TF of *Zm00001d047716*, respectively, indicating the value of the Net-based methods in identifying unique targets. The results of 46 Venn diagrams representing common and unique TGs out of 2,539 pathway genes regulated by the 23 TFs are shown in **Figure D.10**.

To compare the difference in the networks generated by different methods, we showed the networks constructed by the six methods, with the networks constructed by ARANCE method being used as a comparison. All the gene regulatory networks of lignin pathway genes built are shown in the **Figure D.11**. Although each TF's targets were analyzed separately, the results could be merged to obtain a network, in which the TFs

were ranked clockwise based on the number of their connectivities to pathway genes; the TFs with higher connectivities are assumed to regulate more pathway genes and/or have larger impact on pathway genes and thus were ranked earlier. These results indicate that the six methods could be used to rank and select TFs given the TGs are functionally associated structural genes, for example, genes from a pathway or a biological process. In addition, **Figure D.10** manifests that Huber loss function and MSE loss function contribute more to the ranking of TFs than the penalty functions because TFs ranked by Huber-ENET, Huber-Lasso and Huber-Net were more consistent as compared to those by MSE-ENET, MSE-Lasso and MSE-Net.

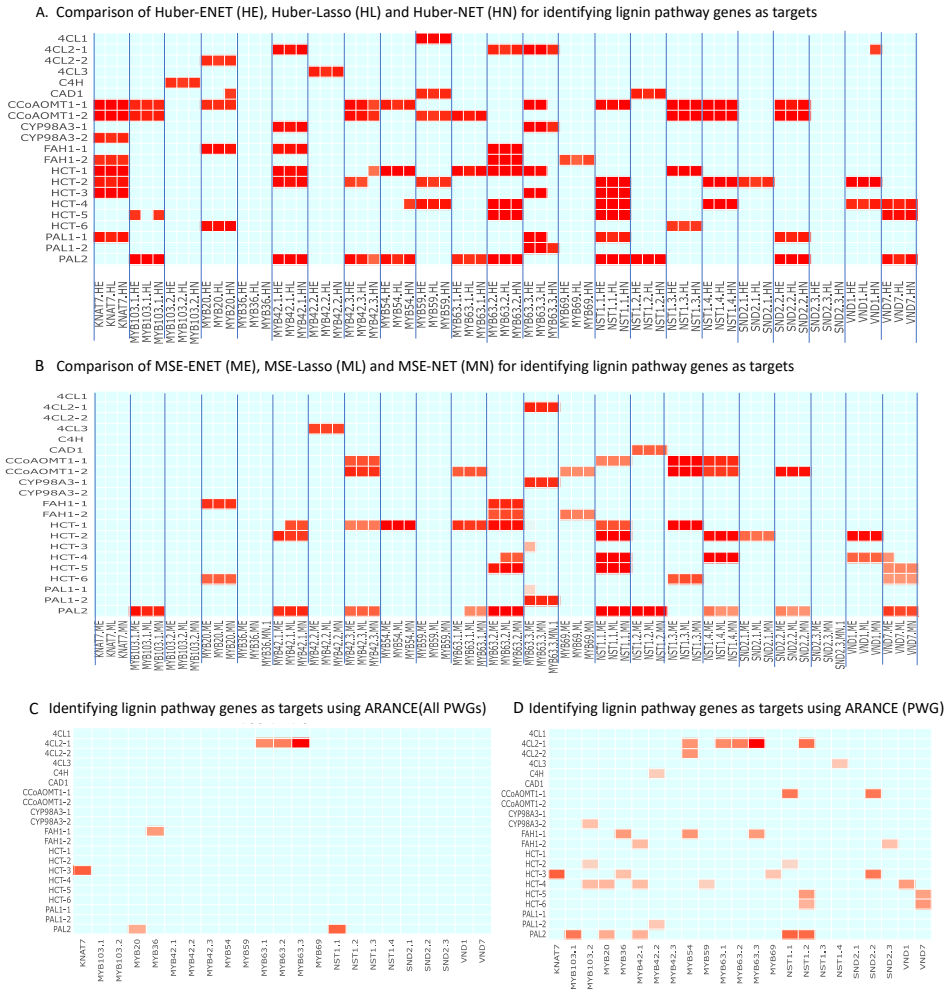


Figure 4.4. Comparison of six methods (Huber-ENET, Huber-Lasso, Huber-Net, MSE-ENET, MSE-Lasso, MSE-Net) in identifying TGs (lignin pathway genes). A. the three methods with Huber loss functions. The inputs are the expression data sets of 2,539 pathway genes and 23 known lignin pathway regulators in the in maize. B. the three methods with MSE loss functions. The inputs are the expression data sets of 2,539 pathway genes and 23 known lignin pathway regulators in the in maize. C. the ARACNE method that was used as a comparison with the same inputs as the six methods. D. the ARACNE method that was used as a comparison with the expression data sets of 21 lignin pathway genes and 23 known lignin pathway regulators being used as the inputs.

4.5 Discussion

4.5.1 Solving Convex optimization problem by implementing APGD

It has been shown that both the loss functions and the penalty functions we used in this study are convex functions[202,220]. Currently, CVX is the commonly used software for solving convex optimization problems[225], but one overt problem of CVX is its slowness when being used for large datasets. In this paper, we implemented an accelerated proximal gradient descent (APGD) algorithm[226] instead of using CVX in our methods. APGD is an effective algorithm to solve an optimization problem with a decomposable objective function, which enabled us to predict true TGs of a TF out of a large number of candidate TGs (e.g. more than 30,000) in the analysis. In principle, CVX cannot be used to calculate the stable selection probability. The stable selection probability is calculated based on the proportion of non-zero estimated regulation effect of a TG over the number of times we resampled in the half-sample approach, and all candidate tuning parameters. When using APGD, we can obtain a subset of TGs with non-zero regulation effects, and the rest subsets of TGs with zero regulation affections. Therefore, we do not need to choose with zero regulation affections. Therefore, we do not need to choose threshold by applying APGD to the half-sample approach.

4.5.2 Development and elucidation of six novel methods for identifying TGs of a TF

With the improved new APGD algorithm, we set out to develop novel methods to predict the TGs of a TF of interest using omics data, an important issue that has not been well solved in current bioinformatics. With two loss functions, Huber and MSE, and three penalty functions, Lasso, ENET and Net, we developed six statistical selection methods, namely, MSE-ENET, Huber-ENET, MSE-Lasso, Huber-Lasso, MSE-Net and Huber-Net. The Huber loss function is a hybrid of squared errors for relatively small errors, and absolute errors for relatively large errors, which has been shown to be more robust than MSE loss function when there are outliers[219]. As the synthetic data generated from the general setting was used to test the first four non-Net methods, we found that MSE-ENET performed better while Huber-Lasso performed worse than other methods if all TGs are independent. When the network setting was used to test the six methods, especially MSE-Net and Huber-Net, as anticipated, the MSE-Net and Huber-Net outperformed the other four non-Net methods since the Net penalty could incorporate the network structure of TGs. Notably, one tuning parameter λ from Lasso penalty and two tuning parameters α and λ from ENET and Net penalties are usually obtained from the cross-validation by minimizing the predicted accuracy[202,230]. However, the results are not stable due to the samples being randomly split in the cross-validation[205]. Therefore, a stability selection method, which uses a subsampling approach to obtain a stable selection result has been developed by Meinshausen and Bühlmann[227]; the subsampling approach has been manifested to determine the amount of regularization. In this study, we used the selection probabilities to evaluate candidate TGs of a given TF.

In our extensive simulation studies, we showed that the proposed methods, Huber-ENET and MSE-ENET, outperformed Huber-Lasso and MSE-Lasso in terms of the true

positive rates. Meanwhile, all of these four methods are useful for predicting or *in-silicon* validating the TGs of a TF of interest in many circumstances. For example, numerous biologists develop transgenic lines or employ a transient system in which a TF is perturbed (**Figure 4.5**), followed by RNA-seq experiments to obtain the transcriptomic data; efficient methods are thus needed to predict or validate the TGs of the TF. Usually, the biologists first identify DEGs after the perturbation of a TF before they use one of the following methods to identify candidate TGs for experimental validation: (1) selecting some DEGs based on significant levels of the corrected p-values or q-values and assume these gene are candidate TGs; (2) using a correlation method as shown[231] or a dependence-based method[232] or a modeling method to identify causal relationships between the TF and DEGs[233]. (3) using Top-down GGM algorithm[198,234,235] to predict TGs of the TF from the DEGs; However, these approaches usually have a low accuracy (e.g. correlation or mutual information) or a scalability limitation (e.g. Top-down GGM) due to the high cost of searching the space of a complete combination of a subset of candidate genes. Thus, there is a pressing need to develop methods for efficient modeling of candidate genes efficiently and predicting the network dynamics accurately. In addition, there are some other circumstances where we need new methods to identify or validate the TGs of a TF *in-silicon*. For example, when genome-wide experiments like ChIP-seq and DAP-seq are conducted, analysis of ChIP-seq or DAP-seq data usually yields a few to even twenty thousand putative TGs whose promoters can be bound by a TF. However, the presence of a binding site of the TF in the TGs' promoters does not necessarily mean there is an activation. We need highly efficient methods to validate the existence of an effect-and-response in expression. In this sense, our methods, Huber-ENET, Huber-Lasso, MSE-ENET and MSE-Lasso, fill in a gap of lacking efficient methods for predicting or validating TGs of a TF of interest using large-scale omics data. Such methods are sought by a multitude of biologists. Our results showed that some TGs identified by our methods couldn't be identified by p-values/FDR-based ranking, Top-down GGM algorithm and correlation/dependence-based methods. Compared to correlation /dependence-based methods that are often applied to pairwise genes, our methods resampled a large number of subsets of data (e.g. 500) to compute the selection probabilities of all genes to one TF simultaneously, and then select top-ranked TGs based on the stabilities of selection probabilities across all subsets. Therefore, our methods augmented the selection process and increased the reliability of TGs. Even if each time we computed linear relationships of one TF with all genomic genes or DEGs with one re-sampled subset, the aggregation to the selection probabilities from all subsets could increase the chance of the nonlinear true relationships to be captured.

Instead of identifying TGs of a TF independently, we sometimes need to investigate if a TF regulates a pathway or a biological process. In this case, we can examine if a TF's TGs contain multiple genes belonging to a pathway or a gene ontology that represents a biology process. Toward this goal, we developed Huber-Net and MSE-Net methods based on network-based penalty. In our extensive simulation studies based on the network setting, we showed that Huber-Net and MSE-Net performed better than the other four methods in terms of the true positive discovery rates. We then applied these two methods to all 2,539 PWGs of maize as candidate TGs and 23 TFs which were identified as the true regulatory TFs of some PWGs in the lignin and phenylpropanoid pathways. By comparing the existing

experimental regulatory relationships from published articles[214-218], the results contained most of the proven positive regulatory relationships. Moreover, we also applied the other four proposed methods, Huber-ENET, MSE-ENET, Huber-Lasso, and MSE-Lasso to examine the differences in the predicted results by two Net methods. We found that most of the regulated relationships are similar while Huber-Net has more rigorous results than others. Thus, the proposed six methods can be used as the reliable methods to predict and/or validate the regulatory relationships between PWGs and TFs.

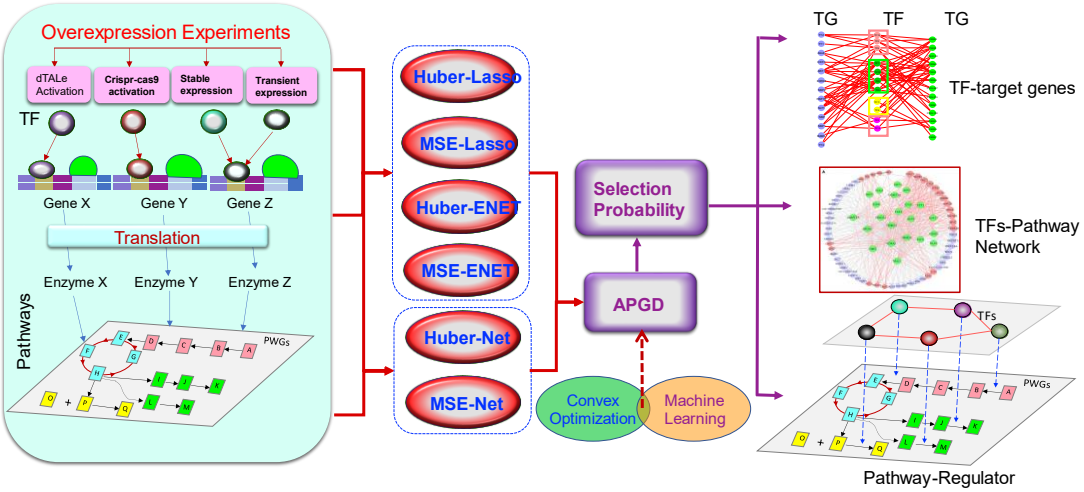


Figure 4.5. An integrative framework for identifying target genes of a TF of interest using transcriptomic data by integration of statistics, machine leaning and convex optimization. Huber and MSE represent Huber loss function and mean squared error MSE, respectively, while ENET, Lasso and Net represent three penalty functions, elastic net, least absolute shrinkage and selection operator**,** and network-based penalty (Net).

4.5.3 The power of statistics, machine learning and optimization combined approaches

In this study, we combined statistics (half-sample approach-derived selection probability), machine learning (regularization in unsupervised learning) and convex optimization (solving regularization with APGD) to identify TGs of a TF of interest, which is illustrated in Figure 5. Our results showed that this kind of combined approach has great efficacy in identifying the true TGs, as we shown early[202].

In our methods, we utilized two loss functions. The Huber loss function is a combination of linear and quadratic loss functions. The MSE loss function, which measures the average of the squared errors, ensures that our trained model has no outlier predictions with huge errors. MSE puts larger weight on these errors due to the squared part of the function. The mathematical benefits of MSE are particularly evident in its use at analyzing the performance of linear regression, as it allows one to partition the variation in a dataset into variation explained by the model and variation explained by randomness. Huber loss is more robust to outliers as compared to MSE loss and least absolute deviation (LAD) loss, and has higher statistical efficiency than the LAD loss function in the absence of outliers[219]. In addition, we utilized three different penalty functions. Lasso penalty adds a

penalty for non-zero coefficients to penalize the sum of their absolute values (l_1 penalty). As a result, for high values of λ, many coefficients are exactly zero under Lasso. ENET penalty was proposed in response to the critique on Lasso because the variable selection of Lasso only considers the absolutely value of estimated effects resulting in instability. It combines the penalties of ridge regression and Lasso to gain "super-penalty". Net penalty is capable of incorporating a set of genes like a pathway or a biological process as represented by a gene ontology and enables us to investigate if a TF regulates multiple genes involved in a pathway or a biology process. When TGs of multiple TFs are predicted, we can then use the results to screen the TFs for regulating a specific metabolic pathway, biological process, and complex trait.

We manifested that APGD has the most computational efficiency for solving the convex optimization problem with both differentiable and undifferentiable functions. Traditional regularization methods need to choose optimal tuning parameters. One limitation of traditional regularization methods with cross-validation is that it depends on the saturation of the data, different data sets may obtain different tuning parameter sets, leading to different or instable results. APGD is a highly efficient approach to solve our proposed methods as well as the other penalized regression, which is a combination of convex optimization and machine learning. The incorporation of half-sample-based selection probability allow to obtain more stable results, and avoid to choose the optimal tuning parameters. Therefore, integration of statistics, machine learning and optimization enables us to take the advantage of all methods and combines them to generate a powerful approach to identify true TGs of a TF with high efficacy. Due to the disadvantage of the feature selection procedure, we cannot check if the selected genes have strong evidence related to the outcome. For future studies, we plan to integrate statistical inference in the selection procedure and further investigate the selection performance by integrating both selection and statistical inference.

4.6 Conclusions

Six new statistical selection methods termed Huber-ENET, MSE-ENET, Huber-Lasso, MSE-Lasso, Huber-Net and MSE-Net were developed for identifying TGs of a TF of interest for the first time by integration of statistics, machine leaning and convex optimization approaches. An accelerated proximal gradient descent algorithm was specifically developed to solve convex optimization. Comprehensive simulations and analyses of the six methods using synthetic data under general setting indicated Huber-ENET, MSE-ENET, Huber-Lasso, and MSE-Lasso could be used to identify true TGs of a TF with high efficacy. When simulating with the data from network setting, Huber-Net and MSE-Net outperformed any other non-Net methods for identifying true TGs involved in a subnetwork. For real data, ENET penalty function appears to contribute greatly to the method efficiency as compared to Lasso, and the Huber optimization has a noticeable contribution to the identification of true TGs of a given TF by increasing the selection probabilities as compared to MSE. AuROC plots showed that all six methods could rank more positive known regulators to the top of output regulatory gene lists. Our results suggest that the overall performances of six methods are instrumental for identifying real TGs of a TF. Our study filled a gap of scarcity of efficient tools for predicting true targets of a TF in gene-wide simulation.

5 Chapter 5

Gene selection by incorporating genetic networks into case-control association studies

Abstract

Large-scale genome-wide association studies (GWAS) have been successfully applied to a wide range of genetic variants underlying complex diseases. The network-based regression approach has been developed to incorporate a biological genetic network and to overcome the challenges caused by the computational efficiency for analyzing high-dimensional genomic data. In this paper, we propose a gene selection approach by incorporating genetic networks into case-control association studies for DNA sequence data or DNA methylation data. Instead of using traditional dimension reduction techniques such as principal component analyses and supervised principal component analyses, we use a linear combination of genotypes at SNPs or methylation values at CpG sites in a gene to capture gene-level signals. We employ three linear combination approaches: optimally weighted sum (OWS), beta-based weighted sum (BWS), and LD-adjusted polygenic risk score (LD-PRS). OWS and LD-PRS are supervised approaches that depend on the effect of each SNP or CpG site on the case-control status, while BWS can be extracted without using the case-control status. After using one of the linear combinations of genotypes or methylation values in each gene to capture gene-level signals, we regularize them to perform gene selection based on the biological network. Simulation studies show that the proposed approaches have higher true positive rates than using traditional dimension reduction techniques. We also apply our approaches to DNA methylation data and UK Biobank DNA sequence data for analyzing rheumatoid arthritis. The results show that the proposed methods can select potentially rheumatoid arthritis related genes that are missed by existing methods.

Keywords: gene selection, genetic network, case-control association study

5.1 Introduction

With the maturation of modern molecular technologies, genomic data is increasingly available in large, diverse data sets[236]. Those data sets provide us an opportunity to use a large volume of human genetic data to explore meaningful insights about diseases. Over the last decade, large-scale genome-wide association studies (GWAS) have been successfully applied to a wide range of genetic variants underlying complex diseases[2]. Different types of genetic variants have different biological functions in the human genome. Genotyping can identify small variations in DNA sequence within populations, such as single-nucleotide polymorphisms (SNPs)[237]. Meanwhile, DNA methylation is an epigenetic marker that has suspected regulatory roles in a broad range of biological processes and diseases[238]. Most penalized regression approaches have been developed to

overcome the challenges caused by the computational efficiency for analyzing high-dimensional genomic data, such as elastic net[239], precision lasso[240], group lasso[241,242], etc. However, Kim et al.[205] showed that these approaches ignore genetic network structures that have the worst selection performance in terms of the true positive rate.

There is strong evidence showing that genes are functionally related to each other in a genetic network and network-based regularization methods by utilizing prior biological network knowledge to select phenotype related genes can outperform other statistical methods that do not utilize genetic network information[205]. Utilizing genetic network information indeed improves selection performance when genomic data are highly correlated among linked genes in the same biological process (i.e., genetic pathway). Therefore, the network-based regularization method has been developed in gene expression data[243] and DNA methylation data[244]. To avoid the computational burden in analyzing high-dimensional genomic data, Kim et al.[205] proposed the approach that combines data dimension reduction techniques with network-based regression to identify phenotype related genes. The dimension reduction techniques can capture the gene-level signals from multiple CpG sites or SNPs in a gene, such as the principal component (PC) based methods (PC, nPC, sPC, et al.)[205]. PC method uses the first PC of DNA methylation data and nPC normalizes the first PC by the largest eigenvalue of the covariance matrix of methylation data. In addition, sPC uses the first PC of the data that only contains the CpG sites associated with the phenotype. It has been demonstrated that network-based regression using PC-based dimension reduction techniques can outperform other methods that ignore genetic network structures[205] and the selection performance can be improved if the gene-level signals can capture more information.

To date, several popular and powerful gene-based association tests for GWAS have been developed to capture the combined effect of individual genetic variants on a phenotype within a gene, including Sequence Kernel Association test (SKAT)[121] and Testing an Optimally Weighted combination of variants (TOW)[50]. The combined effect of individual genetic variants on a phenotype offers an attractive alternative to single genetic variant analysis in GWAS. Let x_{ij} denote the genotype (number of minor alleles) of the i^{th} individual at the j^{th} variant in a gene. To combine information from individual genetic variants into a single measure of risk allele burden, BT, SKAT, and TOW employ a weighted combination of genetic variants, $\sum_j w_j x_{ij}$, to test the association between a gene and a phenotype with different ways to model the weights w_j . SKAT uses the weights related to the minor allele frequencies of the genetic variants. An important feature of SKAT is that it can handle the genetic effects on a phenotype with different directions and magnitudes by incorporating flexible weight functions to boost power. TOW uses the optimal weights obtained by maximizing the score test statistic to test the association between a weighted combination of genetic variants and a phenotype. TOW is more powerful than SKAT when the percentage of neutral variants larger than 50%. However, these three weighted combinations of individual genetic variants do not account for the LD structure among genetic components in a gene. To adjust for LD between genetic variants, the polygenic LD-adjusted risk score (POLARIS) and quadratic polygenic risk score

(PRS_Q) were developed to improve upon the standard PRS by correcting the inflated Type I error rates observed in the standard PRS in the presence of LD[245,246].

Inspired by these popular gene-based association tests using a weighted combination of genetic variants to capture the combined effect of individual genetic variants within a gene, in this paper we propose to use weighted combinations of genetic variants in a gene to capture gene-level signals in network-based regression into case-control association studies with DNA sequence data or DNA methylation data. Instead of using traditional dimension reduction techniques such as PC-based methods, we use a linear combination of genotypes at SNPs or a linear combination of methylation values at CpG sites in each gene to capture gene-level signals. We employ three weighted combinations of variants used in TOW[50], SKAT[121], and PRS_Q[245] to capture gene-level signals. We call these three weighted combinations as optimally weighted sum (OWS), beta-based weighted sum (BWS), and LD-adjusted polygenic risk score (LD-PRS). After we use one of the weighted combinations of genotypes or methylation values in each gene to capture gene-level signals, we regularize them to perform gene selection based on the biological network. Simulation studies show that our proposed methods have higher true positive rates than using traditional dimension reduction techniques. We also apply our methods to DNA methylation data and UK Biobank DNA sequence data for rheumatoid arthritis patients and normal controls. The results show that the methods with the three weighted combinations, OWS, BWS, and LD-PRS, can select potentially rheumatoid arthritis related genes that are missed by the PC-based dimension reduction techniques. Meanwhile, the genes identified by our proposed methods can be significantly enriched into the rheumatoid arthritis pathway, such as genes *HLA-DMA*, *HLA-DPB1*, and *HLA-DQA2* in the HLA region. The overall graphical abstract is summarized in **Figure E.1**.

5.2 Statistical Models and Methods

Consider a sample with n unrelated individuals, indexed by $i = 1, 2, \cdots, n$. Support that there are a set of M genes in the analysis and a total of $\sum_{m=1}^{M} k_m$ genetic components, such as SNPs in DNA sequence data or CpG sites in DNA methylation data, where k_m is the number of genetic components in the m^{th} gene. Let $\mathbf{X}_m = \left(\boldsymbol{x}_1^m, \cdots, \boldsymbol{x}_{k_m}^m \right)$ be an $n \times k_m$ matrix of genetic components in the m^{th} gene, where $\boldsymbol{x}_j^m = \left(x_{1j}^m, \cdots, x_{nj}^m \right)^T$ is the n-dimensional vector which represents the genetic data for the j^{th} genetic component, genotypes of SNPs and M values of CpG sites. Let $\boldsymbol{y} = \left(y_1, \cdots, y_n \right)^T$ be an $n \times 1$ vector of phenotype, where $y_i = 1$ denotes a case and $y_i = 0$ denotes a control in a case-control study. We define a linear combination of genetic components in the m^{th} gene as $\sum_{j=1}^{k_m} w_{mj} \boldsymbol{x}_j^m$.

5.2.1 Weighted linear combination methods

To capture gene-level signals from multiple genetic components in a gene, we employ three weighted combinations of variants, OWS, BWS, and LD-PRS. In the following, we give a summary for each of the weighted combinations. Without loss of generosity, we ignore the

index of the gene and use $\sum_j w_j x_{ij}$ to indicate a linear combination of genetic components in a gene in this section.

OWS uses the weights in TOW to combine the genetic components in a gene. In TOW[50], the weight are determined by maximizing the score test statistic to test the association between $\sum_j w_j x_{ij}$ and a phenotype. The weight are given by $w_j = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_{ij} - \bar{x}_j\right) \Big/ \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$, where $\bar{y}$ and $\bar{x}_j$ represent the sample mean of the phenotype and sample mean of genetic data for the j^{th} genetic component, respectively. Large weight w_j^o represents strong association between the genetic component and the phenotype.

BWS uses the weights given in SKAT[121], where the genetic component is weighted by the beta function, $w_j = \left(Beta\left(\mu_j; a_1, a_2\right)\right)^2$, and is extracted without using the phenotype. For DNA sequence data, $\mu_j = \mathrm{MAF}_j$ and the suggested settings of two parameters in SKAT are $a_1 = 1$ and $a_2 = 25$[121], where MAF_j denotes the minor allele frequency of the j^{th} genetic component in a gene. For DNA methylation data, $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^*$ and x_{ij}^* is the methylation β value for the j^{th} CpG site of the i^{th} individual and $a_1 = a_2 = 0.5$ corresponds to $w_j = 1/\mu_j\left(1-\mu_j\right)$.

Both BWS and OWS are combining the effects of all genetic components in a gene by giving different weights, however, they do not account for LD structure among genetic components in a gene. Motivated by POLARIS[246], we employ the LD-adjusted genetic data to adjust for the influence of LD. The LD-adjusted genetic data is defined as $\tilde{\mathbf{X}} = \mathbf{X}\cdot\mathbf{R}^{-1/2}$, where $\mathbf{R}$ is the correlation matrix of $\mathbf{X}$. However, $\mathbf{R}^{-1/2}$ may not be stable if there are very small eigenvalues of $\mathbf{R}$. To make the LD-adjusted genetic data more robust, we use the method developed by Yan et al.[245] to calculate $\mathbf{R}^{-1/2}$. Let $\lambda_1 \geq \cdots \geq \lambda_k$ and $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_k$ be the eigenvalues and corresponding eigenvectors of $\mathbf{R}$. Then we only use the first J components to calculate $\mathbf{R}^{-1/2}$, where J is the smallest number such that $\sum_{j=1}^{J}\lambda_j \Big/ \sum_{j=1}^{k}\lambda_j \geq 0.999$. Therefore, $\mathbf{R}^{-1/2}$ can be written as $\mathbf{R}^{-1/2} \approx \sum_{j=1}^{J} \boldsymbol{e}_j \boldsymbol{e}_j^T \Big/ \sqrt{\lambda_j}$.

Then LD-PRS uses the weights $w_j = \mathrm{sign}\left(T_j\right)T_j^2$ proposed by Yan et al.[245], where $T_j = \sum_{i=1}^{n} x_{ij}\left(y_i - \bar{y}\right) \Big/ \sqrt{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2 \sum_{i=1}^{n}\left(y_j - \bar{y}\right)^2 / n}$ is the score test statistic to test the association between the j^{th} genetic component and the phenotype. The $\mathrm{sign}\left(T_j\right)$ represents the direction of the effect and T_j^2 represents the strength of the association. Therefore, LD-PRS to capture the gene-level signal is given by $\sum_j w_j \tilde{\boldsymbol{x}}_j$, where $\tilde{\boldsymbol{x}}_j$ is the j^{th} column of $\tilde{\mathbf{X}}$.

Notably, OWS and LD-PRS are supervised methods since their weights are based on the association between each genetic component and the phenotype; BWS is an unsupervised method and the weights depend on the genetic component and not on the phenotype.

5.2.2 Network-based regularization

Consider $\mathbf{A}=(a_{mk})$ is an $M\times M$ adjacency matrix which represents the undirected network connections among genes, where $a_{mk}=1$ represents the m^{th} and k^{th} genes are within the same biological set (i.e., pathway, etc.) and $a_{mk}=0$ otherwise. Let $\mathbf{D}=\text{diag}(d_1,\cdots,d_M)$ be an M dimensional degree matrix, where the m^{th} diagonal element is $d_m=\sum_{k=1}^{M}a_{mk}$ which represents the total number of genetic links of the m^{th} gene. Therefore, the symmetric normalized Laplacian matrix $\mathbf{L}=\mathbf{I}-\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ represents a genetic network structure, where the elements of $\mathbf{L}$ are given by

$$l_{mk}=\begin{cases}1, & \text{if } m=k \text{ and } d_m\neq 0;\\ -(d_m d_k)^{-1/2}, & \text{if } m\neq k,\ d_m\neq 0,\ m \text{ and } k \text{ are linked to each other;}\\ 0, & \text{otherwise.}\end{cases}$$

Let $\mathbf{z}_i=(z_{i1},\cdots,z_{iM})^T$ be a gene-level signal of the i^{th} individual across all genes, which can be obtained by each of the three weighted combinations, OWS, BWS, LD-PRS. Let β_0 and $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_M)^T$ be the intercept and the effect vector of M genes, respectively. The likelihood function of the phenotype is given by

$$L(\beta_0,\boldsymbol{\beta};\mathbf{y})=\prod_{i=1}^{n}f(y_i;\beta_0,\boldsymbol{\beta})=\prod_{i=1}^{n}p(\mathbf{z}_i)^{y_i}\left(1-p(\mathbf{z}_i)\right)^{1-y_i},$$

where $p(\mathbf{z}_i)=\Pr(y_i=1|\mathbf{z}_i)$ represents the probability that the i^{th} individual is a case, which can be calculated by

$$p(\mathbf{z}_i)=\frac{\exp(\beta_0+\mathbf{z}_i^T\boldsymbol{\beta})}{1+\exp(\beta_0+\mathbf{z}_i^T\boldsymbol{\beta})}.$$

Based on the genetic network structure, the penalized logistic likelihood using network-based regularization[205] is given by

$$Q_{\lambda,\alpha}(\beta_0,\boldsymbol{\beta})=-\frac{1}{n}l(\beta_0,\boldsymbol{\beta};\mathbf{y})+P(\boldsymbol{\beta}),$$

where $l(\beta_0,\boldsymbol{\beta};\mathbf{y})=\log L(\beta_0,\boldsymbol{\beta};\mathbf{y})$ is the log-likelihood function and $P(\boldsymbol{\beta})$ is a penalty term which is a combination of the l_1 penalty and squared l_2 penalty incorporating the genetic network structure. $P(\boldsymbol{\beta})$ is defined as

$$P(\boldsymbol{\beta}) = \lambda\alpha\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\mathbf{S}^T\mathbf{L}\mathbf{S}\boldsymbol{\beta} = \lambda\alpha\sum_{m=1}^{M}|\beta_m| + \sum_{m=1}^{M}\sum_{m\sim k}\left(\frac{s_m\beta_m}{\sqrt{d_m}} - \frac{s_k\beta_k}{\sqrt{d_k}}\right)^2,$$

where $\|\bullet\|_1$ is a l_1 norm, and $\mathbf{S} = \text{diag}(s_1, \cdots, s_M)$ is a diagonal matrix of the estimated signs of the regression coefficients on the diagonal entries $s_m \in \{-1, 1\}$ for $m = 1, \cdots, M$, which can be obtained from ordinary regression for $M < n$, and ridge regression for $M \geq n$. λ is a tuning parameter that controls sparsity of the network-based regularization, $\alpha \in [0,1]$ is a mixing proportion between lasso penalty and network-based penalty, and $m \sim k$ denotes that the m^{th} and k^{th} genes are linked to each other in the genetic network.

For a given pair of λ and α, we can estimate the interpret, β_0, and the effect vector of M genes, $\boldsymbol{\beta}$, by minimizing the penalized logistic likelihood $Q_{\lambda,\alpha}(\beta_0, \boldsymbol{\beta})$. It is not difficult to show the penalty function $P(\boldsymbol{\beta})$ is convex [205,224], so the solution β_0 and $\boldsymbol{\beta}$ can be obtained via one of the convex optimization algorithms. We use the R package "pclogit" to estimate β_0 and $\boldsymbol{\beta}$ which implements the cyclic coordinate descent algorithm[244,247]. Same as Chapter 4.2.5, we use half-sample method to calculate the selection probability of each gene, SP_m, for $m = 1, \cdots, M$.

5.3 Simulation Studies

To evaluate if the methods with the three weighted combinations, OWS, LD-PRS, and BWS, outperform the methods with PC-based dimension reduction techniques, we follow the simulation settings in Kim et al.[205] (Details are in **Text E.1**, **Figure E.2**). After generating the individual-level DNA methylation data and DNA sequence data based on a biological network structure, we use the three weighted combinations, OWS, LD-PRS, and BWS, and the three competing PC-based methods, PC, nPC, and sPC, to capture the gene-level signals $\boldsymbol{z}_i = (z_{i1}, \cdots, z_{iM})^T$ for the i^{th} individual across all genes. Then, the selection probability for each gene can be obtained by using a half-sample method 100 times and the network-based regression across 600 pairs of tuning parameters λ and α. We use the true positive rate (TPR) and the area under the receiver operating characteristic (ROC) curve (AUC) to evaluate the selection performance. TPR is defined as the number of true genes that are selected divided by the number of true genes.

For each scenario, we consider a total of $n = 1000$ individuals which contain 500 cases and 500 controls for the balance case-control studies. **Figures 5.1-5.2** show the TPR comparisons for the balance case-control studies in scenario 1. We compare the methods with the three weighted combinations and the methods with the three PC-based dimension reduction techniques, PC, nPC, and sPC, which have been shown higher TPR than other methods that do not utilize biological network information. We first compute selection probabilities of all genes and then rank top genes based on the selection probabilities for each method.

In DNA sequence data analysis (**Figure 5.1** and **Table E.1**), we pre-set the strength of association signals ($\delta = 2,3$), the number of components correlated with the gene-level signal ($\omega = 4,6$), and the error variance which controls the noise level of association signals ($\sigma^2 = 2,3$). The proposed OWS, LD-PRS, and BWS have better selection performance in all eight simulation settings according to TPR and AUC. When the number of causal SNPs in a gene is small ($\omega = 4$), BWS has the uniformly highest TPR and AUC regardless of the size of the error variance. However, selection performance of the supervised approaches, OWS and LD-PRS, are better than or similar as that of the unsupervised approach, BWS, when the number of SNPs in a gene is large ($\omega = 6$). Overall, BWS shows the best selection performance in all simulation settings for DNA sequence data analysis. LD-PRS is better than OWS due to LD-PRS adjusted for the LD structure of the SNPs. In DNA methylation data analysis (**Figure 5.2** and **Table E.2**), we pre-set $\delta = 2, 2.5$, $\omega = 4,6$, and $\sigma^2 = 6,7$. All methods have similar performance according to TPR when the strength of the association signal is small ($\delta = 2$); while the methods with three weighted combinations have higher AUC compared with the three PC-based methods (**Table E.2**). The methods with the three weighted combinations have higher TPRs and AUCs than PC-based methods when the strength of the association signal is large ($\delta = 2.5$). Particularly, when the number of components correlated with the gene-level signal is large ($\omega = 6$), BWS has the uniformly highest TPR regardless of the size of the error variance and the strength of association signals. BWS also shows the best selection performance in all simulation settings for DNA methylation data analysis. LD-PRS and OWS have similar performance but have higher TPRs than the other three PC-based methods.

Figures E.3-E.4 show the TPR comparisons for the balance case-control studies under scenario 2. The patterns of TPR comparisons under scenario 2 for DNA methylation data and DNA sequence data are similar to those under scenario 1 (**Figures 5.1-5.2**). Meanwhile, we also perform TPR comparisons for the unbalance case-control studies, where there are a total of individuals with 100 cases and 900 controls. **Figures E.5-E.8** show the TPR comparisons for the unbalance case-control studies. The patterns of TPR comparisons under these two scenarios for DNA methylation and DNA sequence data are similar to those observed in **Figures 5.1-5.2** and **Figures E.3-E.4**.

We also compare the network-based regression (Net) with two penalized regressions without considering the network structure, elastic net (ENET) and least absolute shrinkage and selection operator (Lasso). The comparison results of the selection performance and the computational time are shown in **Figures E.9-E.13**, which are also explicated in **Text E.2** in more details. In summary, the results show that OWS, LD-PRS, and BWS with Net, always perform better than those with Lasso and ENET. However, three competing PC-based methods (PC, nPC, sPC$n = 1000$) with Net may not increase TPR compared with Lasso and ENET. With respect to model fitting, we use the accuracy rate (ACC) as the measurement for the model fitting quality[248] (**Text E.3**) and we observe that the supervised methods (LD-PRS, OWS, sPC) have higher ACC compared with the three unsupervised methods (BWS, PC, nPC). Notably, LD-PRS and OWS always outperform sPC (**Figure E.14**). Meanwhile, the network-based regression with partially corrected network structure still outperform ENET and Lasso (**Text E.4** and **Figure E.15**).

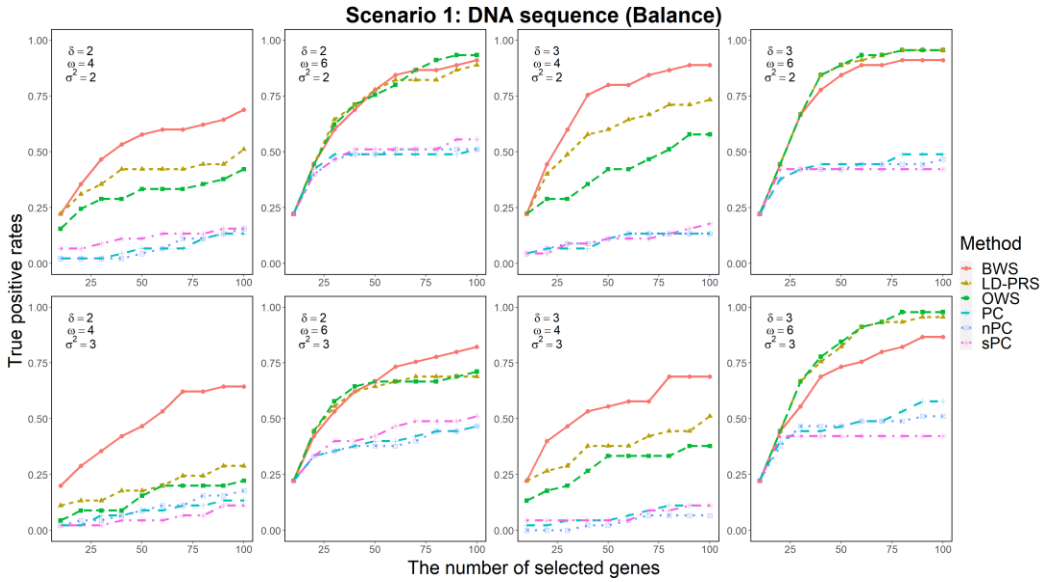


Figure 5.1. The true positive rates of the methods based on different gene-level signals for balance case-control studies with DNA sequence data in scenario 1, where there are five rare variants and five common variants in each gene. According to the different number of selected top genes, three parameters are used to vary the genetic effect: the strength of association signals δ, the number of SNPs in each gene related to gene-level signals ω, and the noise level of association signals σ^2. The selection probabilities are calculated using half-sample method 100 times.

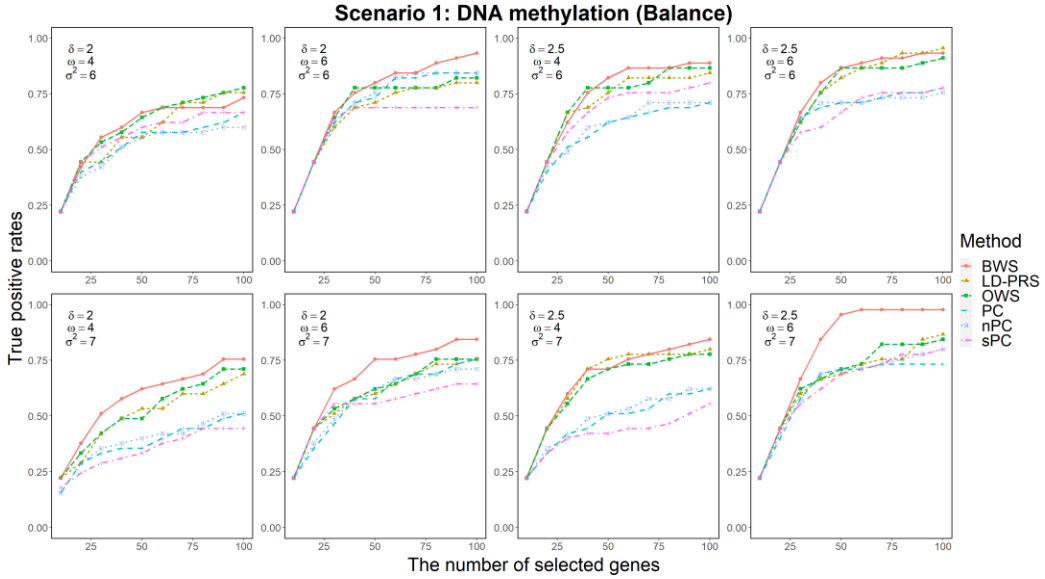


Figure 5.2. The true positive rates of the methods based on different gene-level signals for balance case-control studies with DNA methylation data in scenario 1. According to the different number of selected top-genes, three parameters are used to vary the genetic effect: the strength of association signals δ, the number of CpG sites in each gene related to gene-level signals ω, and the noise level of association signals σ^2. The selection probabilities are calculated using half-sample method 100 times.

5.4 Applications

To evaluate the performance of our proposed methods with three weighted combinations in real data analyses, we apply our methods to DNA methylation data[249,250] and UK Biobank data for DNA sequence of rheumatoid arthritis (RA) patients and normal controls (see details in **Text E.5**). Due to the outperformance of the nPC[205] compared with the other PC-based methods, we only apply nPC to compare the performance with our proposed methods in real data analyses.

5.4.1 Application to DNA methylation data

In the application to DNA methylation data, we select the top 100 genes according to the selection probabilities of each method. We search the GWAS catalog for genes that are associated with RA. **Table 5.1** shows the genes in the GWAS catalog that are also identified by OWS, LD-PRS, BWS, and nPC. OWS identifies 11 genes, LD-PRS identifies 12 genes, BWS identifies 8 genes, and nPC identifies 10 genes. Meanwhile, the number of overlapped genes by each method in the DNA methylation data analysis is summarized in **Figure E.16**. There are four genes identified by all of these four methods, *HLA-DQA2*, *HLA-DRB1*, *HLA-DQB1*, and *CD1C*. Gene *HLA-DRB1*[251] and gene *HLA-DQB1*[67,252-258] play a central role in the immune system and have been reported in the GWAS catalog. No literature reported gene *HLA-DQA2* that was significantly associated with RA in GWAS catalog. However, the SPs of gene *HLA-DQA2* calculated by the methods with the three weighted combinations, OWS, LD-PRS, and BWS, are all 1.000. Also, the SP of gene *HLA-DQA2* is 0.852, which is also on the top 100 genes identified by nPC method. Notably, gene *HLA-DQA2* is in the rheumatoid arthritis pathway (KEGG: hsa05323) and the literature[259] has shown that genes in the human leukocyte antigen (HLA) region remain the most powerful disease risk genes in RA.

Table 5.1. GWAS catalog reported genes identified by OWS, LD-PRS, and BWS in DNA methylation data.

OWS		**LD-PRS**		**BWS**		**nPC**	
Gene	SP	Gene	SP	Gene	SP	Gene	SP
HLA-DRB1	1.000	***HLA-DRB1***	1.000	***HLA-DRB1***	1.000	*CCR6*	1.000
HLA-DRB5	1.000	*KIF26B*	1.000	*PRKCH*	0.998	*ZFP36L1*	1.000
CCR6	0.992	*HLA-DRB5*	0.974	*HLA-DQA1*	0.992	*TCF7*	0.992
ZFP36L1	0.988	*TNXB*	0.974	*HLA-DOB*	0.894	*TNFSF1A*	0.988
NFATC1	0.986	*PRDM16*	0.970	***HLA-DQB1***	0.858	*TLR4*	0.986
TNFRSF1A	0.950	*HLA-DQA1*	0.950	*FNBP1*	0.844	*IL2RB*	0.980
SPSB1	0.928	***HLA-DQB1***	0.950	*TCF7*	0.842	***HLA-DRB1***	0.966
ETS1	0.898	*HLA-DMA*	0.912	*CD247*	0.804	*CD247*	0.962
HLA-DQA1	0.888	*NOTCH4*	0.854			***HLA-DQB1***	0.936
HLA-DQB1	0.880	*HLA-DRA*	0.806			*HLA-DRB5*	0.894
TCF7	0.794	*RIM26*	0.784			*ZNF175*	0.866
		CCR6	0.776				

To better understand the biological meaning behind the top 100 selected genes by each method, we perform the pathway enrichment analysis. In this study, significantly enriched pathways are identified by the top 100 selected genes if FDR < 0.05. In **Figure E.17**, there are 21 significantly enriched pathways identified by OWS, BWS, and LD-PRS, in which the RA pathway is significantly enriched with FDR_{OWS}=1.48E-04, FDR_{BWS} = 7.80E-03, and $FDR_{LD\text{-}PRS}$ = 8.03E-07, respectively; RA pathway is also significantly

enriched in a total of 18 pathways identified by nPC with FDR_{nPC} = 2.91E-03. The overlapping genes between the top 100 genes identified by each method and genes in RA pathway are shown in **Figure 5.3(A)**. The number below each method indicates the total number of overlapping genes identified by the corresponding method and genes in RA pathway. LD-PRS has the smallest pathway enriched FDR and identifies the most overlapping genes (n = 10); genes *HLA-DMA* (SP = 0.912) and *LTB* (SP = 0.998) are uniquely identified. OWS identifies eight overlapping genes which contain one unique gene *HLA-DPB1* (SP = 0.85); meanwhile, BWS identifies six overlapping genes that contain two unique genes *TNF* (SP = 0.980) and *HLA-DOB* (SP = 0.894). Comparing the results of the methods with the three weighted combinations, OWS, LD-PRS, and BWS, and nPC, five HLA-family genes (*HLA-DMA*, *HLA-DOB*, *HLA-DPB1*, *HLA-DPA1*, and *HLA-DQA1*) and two RA pathway genes (*LTB* and *TNF*) are uniquely identified. The results show that the proposed methods can select potentially RA related genes that are missed by nPC.

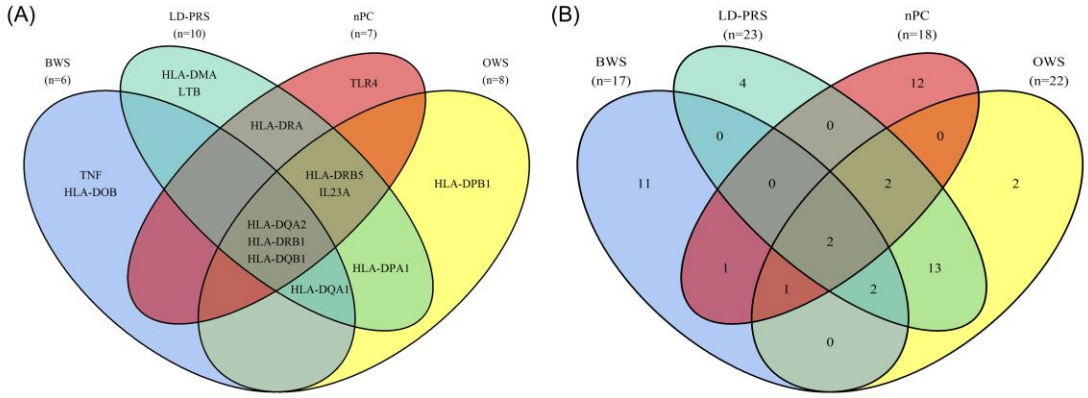


Figure 5.3. Venn diagrams of (A) the number of RA pathway genes identified by BWS, LD-PRS, OWS, and nPC for DNA methylation data; (B) the number of overlapping genes among the top 200 genes identified by each method and reported in the GWAS catalog for DNA sequence data.

5.4.2 Application to DNA sequence data in UK Biobank

In the applications to DNA sequence data, we use 4,541 individuals with RA disease and randomly select 5,459 individuals without RA disease in the UK Biobank. The number of genes with selection probabilities of 1 for DNA sequence data is larger than that of DNA methylation data. For example, there are 80 genes with SP=1 using OWS and 135 genes with SP=1 using LD-PRS. Therefore, we select the top 200 genes according to SPs for DNA sequence data analysis. We also search the GWAS catalog for genes that are associated with RA. **Figure 5.3(B)** and **Table 5.2** show the genes in the GWAS catalog that are also identified by OWS, LD-PRS, BWS, and nPC. Similar to DNA methylation data analyses, LD-PRS identifies the most genes (n=23) reported in the GWAS catalog, including four uniquely identified genes (*HLA-DQB1*, *GFRA1*, *GABBR2*, *EDIL3*); OWS identifies 22 genes in which genes *STAT4* (SP=0.994) and *IKZF1* (SP=0.986) are uniquely selected. There are 13 genes identified by both LD-PRS and OWS, where 12 genes have

selection probabilities of 1 in both methods. Two unsupervised methods, BWS and nPC, can identify 17 and 18 genes in the GWAS catalog. They can uniquely identify 11 and 12 genes, respectively. Moreover, there are two genes identified by all four methods, genes *HLA-DQA1* and *HLA-DRA* (boldfaced in **Table 5.2**), and two genes identified by three proposed methods, genes *RATB* and *CTNNA3*.

Table 5.2. GWAS catalog reported genes identified by OWS, LD-PRS, BWS, and nPC in DNA sequence data.

OWS		**LD-PRS**		**BWS**		**nPC**	
Gene	SP	Gene	SP	Gene	SP	Gene	SP
HLA-DRB1	1.000	*HLA-DRB1*	1.000	***HLA-DRA***	1.000	*HLA-DRB5*	1.000
HLA-DQA1	1.000	***HLA-DQA1***	1.000	***HLA-DQA1***	1.000	***HLA-DQA1***	0.998
PRDM16	1.000	*HLA-DQB1*	1.000	*TNXB*	0.996	*IRF5*	0.966
PRKCB	1.000	***HLA-DRA***	1.000	*HLA-DMA*	0.946	*SOCS2*	0.944
PCSK5	1.000	*PRDM16*	1.000	*SUOX*	0.932	*HLA-DRB1*	0.942
NOTCH4	1.000	*PRKCB*	1.000	*WNT16*	0.930	*TYK2*	0.928
GPC5	1.000	*PCSK5*	1.000	*TYK2*	0.928	*PRDM1*	0.890
RBFOX1	1.000	*NOTCH4*	1.000	*RPS6KB1*	0.902	*NOTCH4*	0.884
DOCK1	1.000	*GPC5*	1.000	*CTNNA3*	0.898	*IL7R*	0.872
KIF26B	1.000	*RBFOX1*	1.000	*HLA-DRB5*	0.892	*ATXN2*	0.872
CTNNA3	1.000	*DOCK1*	1.000	*HIPK1*	0.890	*B3GNT2*	0.870
GALNT18	1.000	*ZMIZ1*	1.000	*SLC9A8*	0.882	*UBE2L3*	0.870
PCDH15	1.000	*SLC9A9*	1.000	*SKIV2L*	0.860	*ELMO1*	0.864
PTPRM	1.000	*RARB*	1.000	*TNIP1*	0.860	*GATA3*	0.846
HLA-DRB5	0.998	*KIF26B*	1.000	*PDF2A*	0.836	*RMI2*	0.844
RARB	0.998	*CTNNA3*	1.000	*TNFAIP3*	0.834	*RORC*	0.836
HLA-DRA	0.996	*GALNT18*	1.000	*RARB*	0.824	***HLA-DRA***	**0.836**
ZMIZ1	0.996	*PCDH15*	1.000			*RBXW8*	0.828
SLC9A9	0.994	*PTPRM*	1.000				
STAT4	0.994	*PDE3A*	0.998				
PDE3A	0.990	*GFRA1*	0.996				
IKZF1	0.986	*GABBR2*	0.994				
		EDIL3	0.992				

Notes: boldface means that the genes are identified by four methods.

5.5 Discussions

In this paper, we employ three weighted combinations to capture the gene-level signals from multiple CpG sites or SNPs: optimally weighted sum (OWS), LD-adjusted polygenic risk score (LD-PRS), and beta-based weighted sum (BWS) in DNA methylation or DNA sequence data. To identify phenotype related genes, we apply the three gene-level signals to a stability gene selection approach by incorporating genetic networks. Compared with the traditional dimension reduction techniques such as PC based gene-level signal, the methods with the three weighted combinations, OWS, LD-PRS, and BWS, have very good performance according to the true positive rates. By applying the methods to real DNA methylation and DNA sequence data, we show that the methods with the three weighted combinations can select more potentially RA related genes that are missed by nPC. Meanwhile, OWS, LD-PRS, and BWS can select more significantly enriched genes in the RA pathway comparing with nPC, such as genes *HLA-DMA*, *HLA-DPB1*, and *HLA-DOB* in the HLA region.

There are some advantages of the three weighted combinations to capture gene-level signals. First, the three weighted combinations can capture more information from genetic components (SNPs or CpG sites) in a gene than the traditional dimension

reduction techniques, such as PC-based methods. OWS and LD-PRS are two supervised approaches based on the association between each genetic component and phenotype, where OWS utilizes the optimally weighted combination[50] of components and LD-PRS can adjust for the highly correlated structure[246] of components. OWS puts large weights on components with large effects on the phenotype[50]. Since the genetic components in a gene are commonly correlated, LD-PRS transforms the original data into an orthogonal space to adjust for LD structure. Moreover, OWS and LD-PRS perform better according to TPR when the genetic components are highly correlated. Even though BWS is an unsupervised method that can be extracted without using phenotype, our simulation studies show that BWS has the highest TPR and AUC in most of the settings. Second, the methods with the three weighted combinations, OWS, LD-PRS, BWS, can select more potential phenotype related genes. In our application to DNA methylation of RA patients and normal controls, the top 100 genes selected by our proposed methods can be significantly enriched into RA pathway and contain more RA pathway genes, especially by LD-PRS. Furthermore, all of our proposed methods have strong evidence to select gene *HLA-QDA2* (SP=1) which is not reported in the GWAS catalog.

Recently, large-scale biobanks linked to electronic health records provide us the possibility of analyzing DNA sequence data using a large sample size. Although three weighted combinations combined with the network-based regression have several advantages, there are three limitations we need to resolve in our future works. First, the method with the three weighted combinations are not suitable for extremely unbalanced case-control studies. To avoid the extremely unbalanced case-control ratio in the data from UK Biobank, we match the number of individuals with and without RA disease in the application of DNA sequence data. This may be the reason for a large number of genes with SP=1 using OWS and LD-PRS, and the SP of the 200th gene using OWS and LD-PRS over 0.97. In the future, we will investigate new methods to handle extremely unbalanced case-control studies. We can use the saddlepoint approximation method[44] to adjust the network-based regression, or use random under-sampling or over-sampling[260] methods instead of using the half-sample approach in the calculation of selection probabilities. The second limitation is that we do not know if the genes selected by the methods with the three weighted combinations are significantly associated with the phenotype. For future studies, we plan to integrate statistical inference in the selection procedure, and further investigate the selection performance by integrating both selection and statistical inference. The third limitation is that the network-based regression is only used for case-control study[205]. For the continuous phenotypes, we need to switch the logistic model with logistic likelihood to the linear regression model with mean squared error or more robust loss function, such as Huber function[261].

6 Reference List

1 Fine, R. S., Pers, T. H., Amariuta, T., Raychaudhuri, S. & Hirschhorn, J. N. Benchmarker: an unbiased, association-data-driven strategy to evaluate gene prioritization algorithms. *The American Journal of Human Genetics* **104**, 1025-1039 (2019).

2 Li, R. *et al.* A regression framework to uncover pleiotropy in large-scale electronic health record data. *Journal of the American Medical Informatics Association* **26**, 1083-1090 (2019).

3 Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**, 5-22 (2017).

4 Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nature Reviews Genetics* **17**, 129-145 (2016).

5 Pendergrass, S. A. *et al.* Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* **9**, e1003087 (2013).

6 Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-wide association studies as a tool to advance precision medicine. *Annual review of genomics and human genetics* **17**, 353-373 (2016).

7 Pendergrass, S. A., Dudek, S. M., Crawford, D. C. & Ritchie, M. D. Visually integrating and exploring high throughput phenome-wide association study (PheWAS) results using PheWAS-view. *BioData mining* **5**, 1-11 (2012).

8 Verma, A. *et al.* Human-disease phenotype map derived from PheWAS across 38,682 individuals. *The American Journal of Human Genetics* **104**, 55-64 (2019).

9 Lee, C. H., Shi, H., Pasaniuc, B., Eskin, E. & Han, B. PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics. *The American Journal of Human Genetics* **108**, 36-48 (2021).

10 Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14**, 483-495 (2013).

11 Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods* **11**, 407-409 (2014).

12 Sha, Q., Wang, Z., Zhang, X. & Zhang, S. A clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. *Bioinformatics* **35**, 1373-1379 (2019).

13 Stephens, M. A unified framework for association analysis with multiple related phenotypes. *PloS one* **8**, e65245 (2013).

14 Yang, Q. & Wang, Y. Methods for analyzing multivariate phenotypes in genetic association studies. *Journal of probability and statistics* **2012** (2012).

15 O'Brien, P. C. Procedures for comparing samples with multiple endpoints. *Biometrics*, 1079-1087 (1984).

16 Liang, X., Wang, Z., Sha, Q. & Zhang, S. An adaptive Fisher's combination method for joint analysis of multiple phenotypes in association studies. *Scientific reports* **6**, 1-10 (2016).

17 Kim, J., Bai, Y. & Pan, W. An adaptive association test for multiple phenotypes with GWAS summary statistics. *Genetic epidemiology* **39**, 651-663 (2015).
18 Yang, J. J., Li, J., Williams, L. K. & Buu, A. An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC bioinformatics* **17**, 1-11 (2016).
19 Cole, D. A., Maxwell, S. E., Arvey, R. & Salas, E. How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological bulletin* **115**, 465 (1994).
20 O'Reilly, P. F. *et al.* MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS one* **7**, e34861 (2012).
21 Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics*, 963-974 (1982).
22 Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22 (1986).
23 Tang, C. S. & Ferreira, M. A. A gene-based test of association using canonical correlation analysis. *Bioinformatics* **28**, 845-850 (2012).
24 Aschard, H. *et al.* Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics* **94**, 662-676 (2014).
25 Wang, Z., Sha, Q. & Zhang, S. Joint analysis of multiple traits using" optimal" maximum heritability test. *PloS one* **11**, e0150975 (2016).
26 Hawkins, R. D., Hon, G. C. & Ren, B. Next-generation genomics: an integrative approach. *Nature Reviews Genetics* **11**, 476-486 (2010).
27 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47**, 1236 (2015).
28 Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* **18**, 117 (2017).
29 O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature genetics* **50**, 1728-1734 (2018).
30 Goh, K.-I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences* **104**, 8685-8690 (2007).
31 Gaynor, S. M., Fagny, M., Lin, X., Platig, J. & Quackenbush, J. Connectivity in eQTL networks dictates reproducibility and genomic properties. *Cell Reports Methods* **2**, 100218 (2022).
32 Tripathi, B., Parthasarathy, S., Sinha, H., Raman, K. & Ravindran, B. Adapting community detection algorithms for disease module identification in heterogeneous biological networks. *Frontiers in genetics* **10**, 164 (2019).
33 Newman, M. *Networks*. (Oxford university press, 2018).
34 Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).
35 Fortunato, S. & Barthelemy, M. Resolution limit in community detection. *Proceedings of the national academy of sciences* **104**, 36-41 (2007).
36 Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Physical review E* **70**, 066111 (2004).

37 Newman, M. E. & Girvan, M. Finding and evaluating community structure in networks. *Physical review E* **69**, 026113 (2004).
38 Newman, M. E. Communities, modules and large-scale structure in networks. *Nature physics* **8**, 25-31 (2012).
39 Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Physics reports* **659**, 1-44 (2016).
40 Barber, M. J. Modularity and community detection in bipartite networks. *Physical Review E* **76**, 066102 (2007).
41 Sudlow, C. *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 med* **12**, e1001779 (2015).
42 Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics* **12**, 417-428 (2011).
43 Pendergrass, S. A. & Crawford, D. C. Using electronic health records to generate phenotypes for research. *Current protocols in human genetics* **100**, e80 (2019).
44 Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *The American Journal of Human Genetics* **101**, 37-49 (2017).
45 Xie, H., Cao, X., Zhang, S. & Sha, Q. Joint analysis of multiple phenotypes for extremely unbalanced case-control association studies. *Genetic Epidemiology*, doi:https://doi.org/10.1002/gepi.22513 (2023).
46 Wang, M., Zhang, S. & Sha, Q. A computationally efficient clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. *PloS one* **17**, e0260911 (2022).
47 Liang, X., Cao, X., Sha, Q. & Zhang, S. HCLC-FC: A novel statistical method for phenome-wide association studies. *Plos one* **17**, e0276646 (2022).
48 Sha, Q., Zhang, Z. & Zhang, S. Joint analysis for genome-wide association studies in family-based designs. *PloS One* **6**, e21957 (2011).
49 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909 (2006).
50 Sha, Q., Wang, X., Wang, X. & Zhang, S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genetic epidemiology* **36**, 561-571 (2012).
51 Nguyen, T., Tagett, R., Diaz, D. & Draghici, S. A novel approach for data integration and disease subtyping. *Genome research* **27**, 2025-2039 (2017).
52 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
53 Bycroft, C. *et al.* Genome-wide genetic data on~ 500,000 UK Biobank participants. *BioRxiv*, 166298 (2017).
54 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-13015-10047-13748 (2015).
55 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1-13 (2009).

56 Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44 (2009).
57 Cao, X., Liang, X., Zhang, S. & Sha, Q. Gene selection by incorporating genetic networks into case-control association studies. *European Journal of Human Genetics*, 1-8 (2022).
58 Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature communications* **8**, 1-11 (2017).
59 de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS computational biology* **11**, e1004219 (2015).
60 Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics* **53**, 1527-1533 (2021).
61 Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics* **99**, 1245-1260 (2016).
62 Tachmazidou, I. *et al.* Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nature genetics* **51**, 230-236 (2019).
63 Kim, S. K., Nguyen, C., Jones, K. B. & Tashjian, R. Z. A Genome Wide Association Study For Shoulder Impingement and Rotator Cuff Disease. *Journal of Shoulder and Elbow Surgery* (2021).
64 Johnston, K. J. *et al.* Genome-wide association study of multisite chronic pain in UK Biobank. *PLoS genetics* **15**, e1008164 (2019).
65 Gorlova, O. *et al.* Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. *PLoS Genet* **7**, e1002178 (2011).
66 Terao, C. *et al.* The human AIRE gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. *Human molecular genetics* **20**, 2680-2685 (2011).
67 Aterido, A. *et al.* Genetic variation at the glycosaminoglycan metabolism pathway contributes to the risk of psoriatic arthritis but not psoriasis. *Annals of the Rheumatic diseases* **78**, 355-364 (2019).
68 Renauer, P. A. *et al.* Identification of susceptibility loci in IL6, RPS9/LILRB3, and an intergenic locus on chromosome 21q22 in Takayasu arteritis in a genome‐wide association study. *Arthritis & rheumatology* **67**, 1361-1368 (2015).
69 Chung, S. A. *et al.* Lupus nephritis susceptibility loci in women with systemic lupus erythematosus. *Journal of the American Society of Nephrology* **25**, 2859-2870 (2014).
70 Cordero, A. I. H. *et al.* Genome-wide associations reveal human-mouse genetic convergence and modifiers of myogenesis, CPNE1 and STC2. *The American Journal of Human Genetics* **105**, 1222-1236 (2019).
71 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).

72 Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology* **9**, 770-780 (2008).
73 Sharma, A. *et al.* Network-based analysis of genome wide association data provides novel candidate genes for lipid and lipoprotein traits. *Molecular & Cellular Proteomics* **12**, 3398-3408 (2013).
74 Vinayagam, A. *et al.* Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Sciences* **113**, 4976-4981, doi:doi:10.1073/pnas.1603992113 (2016).
75 Loscalzo, J. *Network medicine*. (Harvard University Press, 2017).
76 Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature reviews genetics* **12**, 56-68 (2011).
77 Cao, X., Zhang, S. & Sha, Q. A novel method for multiple phenotype association studies based on genotype and phenotype network. *bioRxiv*, 2023.2002. 2023.529687 (2023).
78 Abdellaoui, A., Yengo, L., Verweij, K. J. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics* (2023).
79 Korte, A. & Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* **9**, 1-9 (2013).
80 Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews genetics* **5**, 101-113 (2004).
81 Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *science* **286**, 509-512 (1999).
82 Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17-60 (1960).
83 Borgatti, S. P. Centrality and network flow. *Social networks* **27**, 55-71 (2005).
84 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47**, 1228-1235 (2015).
85 Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics* **49**, 1421-1427 (2017).
86 Kim, S. S. *et al.* Genes with high network connectivity are enriched for disease heritability. *The American Journal of Human Genetics* **104**, 896-913 (2019).
87 Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proceedings of the national academy of sciences* **99**, 7821-7826 (2002).
88 Fortunato, S. Community detection in graphs. *Physics reports* **486**, 75-174 (2010).
89 Liu, Z. & Lin, X. Multiple phenotype association tests using summary statistics in genome‐wide association studies. *Biometrics* **74**, 165-175 (2018).
90 Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 479-498 (2002).
91 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440-9445 (2003).

92 Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association* **96**, 1151-1160 (2001).
93 Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *The annals of statistics* **31**, 2013-2035 (2003).
94 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289-300 (1995).
95 Cao, X., Shi, Y., Wang, P., Chen, L. & Wang, Y. in *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*. 329-333 (IEEE).
96 Easley, D. & Kleinberg, J. *Networks, crowds, and markets: Reasoning about a highly connected world*. (Cambridge university press, 2010).
97 Newman, M. E. The structure and function of complex networks. *SIAM review* **45**, 167-256 (2003).
98 Latapy, M., Magnien, C. & Del Vecchio, N. Basic notions for the analysis of large two-mode networks. *Social networks* **30**, 31-48 (2008).
99 Kullback, S. & Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics* **22**, 79-86 (1951).
100 Murphy, K. P. *Machine learning: a probabilistic perspective*. (MIT press, 2012).
101 Dey, K. K. *et al.* SNP-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. *Cell genomics* **2**, 100145 (2022).
102 Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics* **50**, 621-629 (2018).
103 Xie, H., Cao, X., Zhang, S. & Sha, Q. Joint analysis of multiple phenotypes for extremely unbalanced case-control association studies using multi-layer network. *submitted* (2023+).
104 Kim, Y., Son, S.-W. & Jeong, H. Finding communities in directed networks. *Physical Review E* **81**, 016103 (2010).
105 Mikhail, D., Anton, K. & Denis, T. Parallel modularity computation for directed weighted graphs with overlapping communities. *Труды Института системного программирования РАН* **28**, 153-170 (2016).
106 Liu, Y. *et al.* ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics* **104**, 410-421 (2019).
107 Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics* **50**, 229-237 (2018).
108 Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *The American Journal of Human Genetics* **96**, 21-36 (2015).
109 Zhang, Y. *et al.* SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome biology* **22**, 1-30 (2021).

110 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47**, 1236-1241 (2015).
111 Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics* **88**, 586-598 (2011).
112 Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics* **95**, 5-23 (2014).
113 Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179-189 (2020).
114 Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS genetics* **17**, e1009440 (2021).
115 Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genetics* **18**, e1010299 (2022).
116 Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48**, 245-252 (2016).
117 Hebbring, S. J. The challenges, advantages and future of phenome‐wide association studies. *Immunology* **141**, 157-165 (2014).
118 Kraft, P., Zeggini, E. & Ioannidis, J. P. Replication in genome-wide association studies. *Statistical science: a review journal of the Institute of Mathematical Statistics* **24**, 561 (2009).
119 Li, M.-X., Gui, H.-S., Kwan, J. S. & Sham, P. C. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *The American Journal of Human Genetics* **88**, 283-293 (2011).
120 Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311-321 (2008).
121 Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82-93 (2011).
122 Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762-775 (2012).
123 Svishcheva, G. R., Belonogova, N. M., Zorkoltseva, I. V., Kirichenko, A. V. & Axenovich, T. I. Gene-based association tests using GWAS summary statistics. *Bioinformatics* **35**, 3701-3708 (2019).
124 Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nature reviews genetics* **18**, 117-127 (2017).
125 Conneely, K. N. & Boehnke, M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *The American Journal of Human Genetics* **81**, 1158-1168 (2007).
126 Kwak, I.-Y. & Pan, W. Adaptive gene-and pathway-trait association testing with GWAS summary statistics. *Bioinformatics* **32**, 1178-1184 (2016).
127 de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015).

128 Zhang, J., Xie, S., Gonzales, S., Liu, J. & Wang, X. A fast and powerful eQTL weighted method to detect genes associated with complex trait using GWAS summary data. *Genetic epidemiology* **44**, 550-563 (2020).
129 Consortium, G. P. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
130 Deng, Y. & Pan, W. Improved use of small reference panels for conditional and joint analysis with GWAS summary statistics. *Genetics* **209**, 401-408 (2018).
131 Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics* **99**, 139-153 (2016).
132 Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906-2914 (2014).
133 Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47**, 1091 (2015).
134 Xu, Z., Wu, C., Wei, P. & Pan, W. A powerful framework for integrating eQTL and GWAS summary data. *Genetics* **207**, 893-902 (2017).
135 Van der Sluis, S., Posthuma, D. & Dolan, C. V. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet* **9**, e1003235 (2013).
136 Dutta, D. *et al.* A powerful subset-based method identifies gene set associations and improves interpretation in UK Biobank. *The American Journal of Human Genetics* **108**, 669-681 (2021).
137 Wu, C. Multi-trait Genome-Wide Analyses of the Brain Imaging Phenotypes in UK Biobank. *Genetics* **215**, 947-958, doi:10.1534/genetics.120.303242 (2020).
138 Yang, Y., Basu, S., Mirabello, L., Spector, L. & Zhang, L. A Bayesian gene-based genome-wide association study analysis of osteosarcoma trio data using a hierarchically structured prior. *Cancer informatics* **17**, 1176935118775103 (2018).
139 Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *The American Journal of Human Genetics* **93**, 42-53 (2013).
140 Hogg, R. V., Tanis, E. A. & Zimmerman, D. L. *Probability and statistical inference*. Vol. 993 (Macmillan New York, 1977).
141 Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nature genetics* **46**, 430-437 (2014).
142 Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature communications* **9**, 1-20 (2018).
143 Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association* **115**, 393-402 (2020).
144 Rosseel, Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software* **48**, 1-36 (2012).

145 Nagpal, S. *et al.* TIGAR: an improved Bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *The American Journal of Human Genetics* **105**, 258-266 (2019).
146 Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics* **45**, 1150 (2013).
147 Consortium, S. W. G. o. t. P. G. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
148 Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics* **50**, 381-389 (2018).
149 Ikeda, M. *et al.* Genome-wide association study detected novel susceptibility genes for schizophrenia and shared trans-populations/diseases genetic effect. *Schizophrenia bulletin* **45**, 824-834 (2019).
150 Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature genetics* **49**, 1576 (2017).
151 Goes, F. S. *et al.* Genome‐wide association study of schizophrenia in Ashkenazi Jews. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **168**, 649-659 (2015).
152 Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nature genetics* **51**, 1670-1678 (2019).
153 Periyasamy, S. *et al.* Association of schizophrenia risk with disordered niacin metabolism in an Indian genome-wide association study. *JAMA psychiatry* **76**, 1026-1034 (2019).
154 Lee, P. H. *et al.* Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469-1482. e1411 (2019).
155 Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Molecular autism* **8**, 1-17 (2017).
156 Lam, M. *et al.* Pleiotropic meta-analysis of cognition, education, and schizophrenia differentiates roles of early neurodevelopmental and adult synaptic pathways. *The American Journal of Human Genetics* **105**, 334-350 (2019).
157 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713 (2010).
158 Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature genetics* **45**, 1274 (2013).
159 Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769-773 (2018).
160 Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nature genetics* **43**, 969 (2011).
161 Kilpeläinen, T. O. *et al.* Multi-ancestry study of blood lipid levels identifies four loci interacting with physical activity. *Nature communications* **10**, 1-11 (2019).
162 Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514-518 (2019).
163 Liu, H. *et al.* Heritability and genome-wide association study of plasma cholesterol in chinese adult twins. *Frontiers in endocrinology* **9**, 677 (2018).

164 Spracklen, C. N. *et al.* Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels. *Human molecular genetics* **26**, 1770-1784 (2017).
165 De Vries, P. S. *et al.* Multiancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions. *American journal of epidemiology* **188**, 1033-1054 (2019).
166 Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nature genetics* **50**, 401-413 (2018).
167 Ripatti, P. *et al.* Polygenic Hyperlipidemias and Coronary Artery Disease Risk. *Circulation: Genomic and Precision Medicine* **13**, e002725 (2020).
168 Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS medicine* **17**, e1003062 (2020).
169 Noordam, R. *et al.* Multi-ancestry sleep-by-SNP interaction analysis in 126,926 individuals reveals lipid loci stratified by sleep duration. *Nature communications* **10**, 1-13 (2019).
170 Klarin, D. *et al.* Genetics of blood lipids among~ 300,000 multi-ethnic participants of the Million Veteran Program. *Nature genetics* **50**, 1514-1523 (2018).
171 Qi, G. & Chatterjee, N. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. *PLoS genetics* **14**, e1007549 (2018).
172 Klimentidis, Y. C. *et al.* Phenotypic and genetic characterization of lower LDL cholesterol and increased type 2 diabetes risk in the UK Biobank. *Diabetes* **69**, 2194-2205 (2020).
173 Liu, D. J. *et al.* Exome-wide association study of plasma lipids in> 300,000 individuals. *Nature genetics* **49**, 1758-1766 (2017).
174 Curtis, D., Vine, A. E. & Knight, J. A simple method for assessing the strength of evidence for association at the level of the whole gene. *Advances and applications in bioinformatics and chemistry: AABC* **1**, 115 (2008).
175 Wang, M. *et al.* COMBAT: a combined association test for genes using summary statistics. *Genetics* **207**, 883-891 (2017).
176 Gerring, Z. F., Mina-Vargas, A., Gamazon, E. R. & Derks, E. M. E-MAGMA: an eQTL-informed method to identify risk genes using genome-wide association study summary statistics. *Bioinformatics* **37**, 2245-2249 (2021).
177 Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics* **48**, 481-487 (2016).
178 Feng, H. *et al.* Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. *PLoS genetics* **17**, e1008973 (2021).
179 Zhu, H., Zhang, S. & Sha, Q. A novel method to test associations between a weighted combination of phenotypes and genetic variants. *PloS one* **13**, e0190788 (2018).

180 Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *The Lancet* **361**, 598-604 (2003).
181 Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nature genetics* **36**, 388-393 (2004).
182 Chen, T., He, H. L. & Church, G. M. Modeling gene expression with differential equations. *Biocomputing'99*, 29-40 (1999).
183 Ruklisa, D., Brazma, A. & Viksna, J. Reconstruction of gene regulatory networks under the finite state linear model. *Genome informatics* **16**, 225-236 (2005).
184 Dojer, N., Gambin, A., Mizera, A., Wilczyński, B. & Tiuryn, J. Applying dynamic Bayesian networks to perturbed gene expression data. *BMC bioinformatics* **7**, 1-11 (2006).
185 Kauffman, S. Homeostasis and differentiation in random genetic control networks. *Nature* **224**, 177-178 (1969).
186 Chen, B.-S., Chang, C.-H., Wang, Y.-C., Wu, C.-H. & Lee, H.-C. Robust model matching design methodology for a stochastic synthetic gene network. *Mathematical biosciences* **230**, 23-36 (2011).
187 Cao, J., Qi, X. & Zhao, H. Modeling gene regulation networks using ordinary differential equations. *Next generation microarray bioinformatics*, 185-197 (2012).
188 De La Fuente, A., Bing, N., Hoeschele, I. & Mendes, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**, 3565-3574 (2004).
189 Meyer, P. E., Lafitte, F. & Bontempi, G. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics* **9**, 1-10 (2008).
190 Butte, A. J. & Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Biocomputing 2000*, 418-429 (1999).
191 Margolin, A. A. *et al.* in *BMC bioinformatics.* 1-15 (BioMed Central).
192 Faith, J. J. *et al.* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology* **5**, e8 (2007).
193 Altay, G. & Emmert-Streib, F. Structural influence of gene networks on their inference: analysis of C3NET. *Biology Direct* **6**, 1-16 (2011).
194 Luo, W., Hankenson, K. D. & Woolf, P. J. Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. *BMC bioinformatics* **9**, 1-15 (2008).
195 Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *Journal of computational biology* **7**, 601-620 (2000).
196 Huynh-Thu, V. A. & Geurts, P. Unsupervised gene network inference with decision trees and random forests. *Gene Regulatory Networks*, 195-215 (2019).
197 Lin, Y.-C. *et al.* SND1 transcription factor–directed quantitative functional hierarchical genetic regulatory network in wood formation in Populus trichocarpa. *The Plant Cell* **25**, 4324-4341 (2013).

198 Wei, H. Construction of a hierarchical gene regulatory network centered around a transcription factor. *Briefings in Bioinformatics* **20**, 1021-1031 (2019).
199 Kumari, S. *et al.* Bottom-up GGM algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways or processes. *BMC bioinformatics* **17**, 1-12 (2016).
200 Deng, W., Zhang, K., Busov, V. & Wei, H. Recursive random forest algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways. *PloS one* **12**, e0171532 (2017).
201 Gunasekara, C., Zhang, K., Deng, W., Brown, L. & Wei, H. TGMI: an efficient algorithm for identifying pathway regulators through evaluation of triple-gene mutual interaction. *Nucleic acids research* **46**, e67-e67 (2018).
202 Deng, W., Zhang, K., He, C., Liu, S. & Wei, H. HB-PLS: A statistical method for identifying biological process or pathway regulators by integrating Huber loss and Berhu penalty with partial least squares regression. *Forestry Research* **1**, 1-13 (2021).
203 Danaher, P., Wang, P. & Witten, D. M. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 373-397 (2014).
204 Deng, W. *et al.* JRmGRN: joint reconstruction of multiple gene regulatory networks with common hub genes using data from multiple tissues or conditions. *Bioinformatics* **34**, 3470-3478 (2018).
205 Kim, K. & Sun, H. Incorporating genetic networks into case-control association studies with high-dimensional DNA methylation data. *BMC bioinformatics* **20**, 1-15 (2019).
206 Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics* **74**, 47 (2002).
207 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics* **26**, 139-140 (2010).
208 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
209 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
210 Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology* **12**, 1-14 (2011).
211 Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology* **31**, 46-53 (2013).
212 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515 (2010).
213 Hawkins, C. *et al.* Plant Metabolic Network 15: A resource of genome‐wide metabolism databases for 126 plants and algae. *Journal of integrative plant biology* **63**, 1888-1905 (2021).

214 Han, X. *et al.* Lignin biosynthesis and accumulation in response to abiotic stresses in woody plants. *Forestry Research* **2** (2022).
215 Liu, Y. *et al.* Functional characterization of Populus PsnSHN2 in coordinated regulation of secondary wall components in tobacco. *Scientific reports* **7**, 1-11 (2017).
216 Xie, M. *et al.* Regulation of lignin biosynthesis and its role in growth-defense tradeoffs. *Frontiers in plant science* **9**, 1427 (2018).
217 Zhong, R., Lee, C., Zhou, J., McCarthy, R. L. & Ye, Z.-H. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. *The Plant Cell* **20**, 2763-2782 (2008).
218 Zhong, R. & Ye, Z.-H. MYB46 and MYB83 bind to the SMRE sites and directly activate a suite of transcription factors and secondary wall biosynthetic genes. *Plant and Cell Physiology* **53**, 368-380 (2012).
219 Owen, A. B. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics* **443**, 59-72 (2007).
220 Cao, X., Liang, X., Zhang, S. & Sha, Q. Gene selection by incorporating genetic networks into case-control association studies. *bioRxiv* (2022).
221 Ihmels, J., Levy, R. & Barkai, N. Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nature biotechnology* **22**, 86-92 (2004).
222 Wei, H. *et al.* Transcriptional coordination of the metabolic network in Arabidopsis. *Plant physiology* **142**, 762-774 (2006).
223 Li, C. & Li, H. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics* **4**, 1498 (2010).
224 Choi, J., Kim, K. & Sun, H. New variable selection strategy for analysis of high-dimensional DNA methylation data. *Journal of bioinformatics and computational biology* **16**, 1850010 (2018).
225 Grant, M., Boyd, S. & Ye, Y. (2008).
226 Parikh, N. & Boyd, S. Proximal algorithms. *Foundations and trends® in Optimization* **1**, 127-239 (2014).
227 Meinshausen, N. & Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417-473 (2010).
228 Zhao, M. *et al.* Bacterium-Enabled Transient Gene Activation by Artificial Transcription Factor for Resolving Gene Regulation in Maize. *bioRxiv* (2021).
229 (!!! INVALID CITATION !!! 144-148).
230 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67**, 301-320 (2005).
231 Kumari, S. *et al.* Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PloS one* **7**, e50411 (2012).
232 Akhand, M., Nandi, R., Amran, S. & Murase, K. in *2015 18th International Conference on Computer and Information Technology (ICCIT).* 312-316 (IEEE).

233 Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat Commun* **9**, 1090, doi:10.1038/s41467-018-03424-4 (2018).
234 Wei, M. *et al.* PuHox52 - mediated hierarchical multilayered gene regulatory network promotes adventitious root formation in Populus ussuriensis. *New Phytologist* **228**, 1369-1385 (2020).
235 Wu, W. *et al.* Growth - regulating factor 5 (GRF5) - mediated gene regulatory network promotes leaf growth and expansion in poplar. *New Phytologist* **230**, 612-628 (2021).
236 Ritchie, M. D. Large-scale analysis of genetic and clinical patient data. *Annual Review of Biomedical Data Science* **1**, 263-274 (2018).
237 Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077-1082 (1998).
238 Bock, C. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics* **13**, 705-719 (2012).
239 Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C. & Sölkner, J. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics* **4**, 270 (2013).
240 Wang, H., Lengerich, B. J., Aragam, B. & Xing, E. P. Precision Lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics* **35**, 1181-1187 (2019).
241 Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49-67 (2006).
242 Meier, L., Van De Geer, S. & Bühlmann, P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 53-71 (2008).
243 Li, C. & Li, H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**, 1175-1182 (2008).
244 Sun, H. & Wang, S. Network - based regularization for matched case - control analysis of high - dimensional DNA methylation data. *Statistics in medicine* **32**, 2127-2139 (2013).
245 Yan, S., Sha, Q. & Zhang, S. Gene-Based Association Tests Using New Polygenic Risk Scores and Incorporating Gene Expression Data. *Genes* **13**, 1120 (2022).
246 Baker, E. *et al.* POLARIS: Polygenic LD - adjusted risk score approach for set - based analysis of GWAS data. *Genetic epidemiology* **42**, 366-377 (2018).
247 Sun, H. & Wang, S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics* **28**, 1368-1375 (2012).
248 Kuhn, M. & Johnson, K. *Applied predictive modeling*. Vol. 26 (Springer, 2013).
249 Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology* **31**, 142-147 (2013).
250 Kular, L. *et al.* DNA methylation as a mediator of HLA-DRB1* 15: 01 and a protective variant in multiple sclerosis. *Nature communications* **9**, 1-15 (2018).

251 (!!! INVALID CITATION !!! 224-230).
252 Jiang, X. *et al.* An Immunochip-based interaction study of contrasting interaction effects with smoking in ACPA-positive versus ACPA-negative rheumatoid arthritis. *Rheumatology* **55**, 149-155 (2016).
253 Wei, W.-H., Viatte, S., Merriman, T. R., Barton, A. & Worthington, J. Genotypic variability based association identifies novel non-additive loci DHCR7 and IRF4 in sero-negative rheumatoid arthritis. *Scientific reports* **7**, 1-7 (2017).
254 Julia, A. *et al.* Genome‐wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* **58**, 2275-2286 (2008).
255 Negi, S. *et al.* A genome‐wide association study reveals ARL15, a novel non‐HLA susceptibility gene for rheumatoid arthritis in North Indians. *Arthritis & Rheumatism* **65**, 3026-3035 (2013).
256 Kochi, Y. *et al.* A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. *Nature genetics* **42**, 515-519 (2010).
257 Raychaudhuri, S. *et al.* Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nature genetics* **40**, 1216-1223 (2008).
258 Consortium, W. T. C. C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 (2007).
259 Weyand, C. M. & Goronzy, J. J. Association of MHC and rheumatoid arthritis: HLA polymorphisms in phenotypic variants of rheumatoid arthritis. *Arthritis Research & Therapy* **2**, 1-5 (2000).
260 Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321-357 (2002).
261 Huber, P. J. in *Breakthroughs in statistics* 492-518 (Springer, 1992).

A Supplementary Materials for Chapter 1

A.1 Supplementary Text

Text A.1. Details of the six multiple phenotype association tests.

To test the association between phenotypes in each network module and a SNP, we perform the following six multiple phenotypes association tests. To simplify the notation, we assume that the tests are applied to test the association between K phenotypes and a SNP.

CLC[1]: CLC classifies K phenotypes into L clusters for $L=1,\cdots,K$. The test statistic with L clusters is $T_{CLC}^{L}=\left(\mathbf{D}\boldsymbol{T}\right)^{T}\left(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^{T}\right)^{-1}\left(\mathbf{D}\boldsymbol{T}\right)$, where $\boldsymbol{T}=\left(T_1,\cdots,T_K\right)^{T}$ with T_k being the score test statistic to test the association between the k^{th} phenotype and a SNP; $\mathbf{D}=\mathbf{B}^{T}\boldsymbol{\Sigma}^{-1}$ with $\boldsymbol{\Sigma}$ being a correlation matrix of K phenotypes and $\mathbf{B}=\left(b_{kl}\right)$ being a $K\times L$ matrix with $b_{kl}=1$ if the k^{th} phenotype belongs to the l^{th} cluster and $b_{kl}=0$ otherwise. Under the null hypothesis that none of the K phenotypes are associated with a SNP, T_{CLC}^{L} follows a chi-square distribution with degrees of freedom L. The overall test statistic of CLC is given by $T_{CLC}=\max_{1\le L\le K}T_{CLC}^{L}$ and the corresponding p-value can be evaluated by a simulation procedure.

ceCLC[2]: ceCLC is a computational efficient version of CLC, where the p-value of overall test statistic is derived by the Cauchy combination method[3,4]. Let p_L be the p-value of T_{CLC}^{L}, then the test statistic of ceCLC is given by $T_{ceCLC}=\sum_{L=1}^{K}\tan\left\{\left(0.5-p_L\right)\pi\right\}\big/K$. The null distribution of T_{ceCLC} can be well approximated by a standard Cauchy distribution. Therefore, p-value of ceCLC can be approximated by $p_{ceCLC}=0.5-\left\{\arctan\left(T_{ceCLC}\right)/\pi\right\}$.

HCLC[5]: Instead of considering all possible number of clusters in CLC and ceCLC, HCLC determine the optimal number of clusters, L^{*}, by using a stopping criterion that maximizes the cluster separation[6]. Therefore, the test statistic of HCLC is defined as $T_{HCLC}=T_{CLC}^{L^{*}}$ and the p-value is calculated by assuming T_{HCLC} follows a chi-square distribution with degrees of freedom L^{*}.

MultiPhen[7]: MultiPhen uses the ordinal regression (also known as proportional odds logistic regression) to regress genotype of a SNP on K phenotypes. MultiPhen uses a likelihood ratio test to test whether effect sizes of K phenotypes are significantly different from zero. The resulting test statistic asymptotically follows a chi-square distribution with degrees of freedom . K ..

O'Brien[8]: O'Brien uses a linear combination method of the score test statistic, T_k , to test the association between the k^{th} phenotype and a SNP. That is, the test statistic of O'Brien is given by $T_{O'Brien}=\left(\boldsymbol{1}_K^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{T}\right)^{2}$, where $\boldsymbol{1}_K$ is a $K\times 1$ vector with elements of all 1s. Under the null hypothesis, $T_{O'Brien}$ follows a chi-square distribution with 1 degree of freedom.

Omnibus[1]: Omnibus is developed to overcome the limitation of O'Brien. The test statistic of Omnibus is $T_{omnibus} = \boldsymbol{T}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{T}$. Under the null hypothesis, $T_{omnibus}$ follows a chi-square distribution with degrees of freedom K.

Note that a normal approximation of the score test statistic T_k for $k = 1, \cdots, K$ used in CLC, ceCLC, HCLC, O'Brien, and Omnibus has inflated type I error rates from binary phenotypes with extremely unbalanced case-control ratios [9]. In this case, we modify these five methods by calculating $T_k = \hat{\beta}_k / se\left(\hat{\beta}_k\right)$, where $\hat{\beta}_k$ and $se\left(\hat{\beta}_k\right)$ can be estimated by saddlepoint approximation[9].

A.2 Supplementary Tables

Table A.1. Simulation settings with $\hat{\boldsymbol{\lambda}}_1 = \beta(1, \cdots, 1)^T$ and $\hat{\boldsymbol{\lambda}}_2 = \frac{2\beta}{k+1}(1, \cdots, k)^T$.

	Category	SNP 1-50	SNP 51-100	SNP 101-150	SNP 151-200
Model 1	1	$\hat{\boldsymbol{\lambda}}_1$	0	0	0
	2	0	$-\hat{\boldsymbol{\lambda}}_1$	0	0
Model 2	1	$\hat{\boldsymbol{\lambda}}_1$	0	0	0
	2	0	$-\hat{\boldsymbol{\lambda}}_2$	0	0
Model 3	1	$\hat{\boldsymbol{\lambda}}_1$	0	0	0
	2	0	$-\hat{\boldsymbol{\lambda}}_1$	0	0
	3-5	0	0	0	0
Model 4	1	$\hat{\boldsymbol{\lambda}}_1$	0	0	0
	2	0	$-\hat{\boldsymbol{\lambda}}_2$	0	0
	3-5	0	0	0	0
Model 5	1	$\hat{\boldsymbol{\lambda}}_1$	0	0	0
	2	0	$-\hat{\boldsymbol{\lambda}}_1$	0	0
	3	0	0	$\hat{\boldsymbol{\lambda}}_2$	0
	4	0	0	0	$-\hat{\boldsymbol{\lambda}}_2$
Model 6	1	$\hat{\boldsymbol{\lambda}}_1$	0	0	0
	2	0	$-\hat{\boldsymbol{\lambda}}_1$	0	0
	3	0	0	$\hat{\boldsymbol{\lambda}}_2$	0
	4	0	0	0	$-\hat{\boldsymbol{\lambda}}_2$
	5-10	0	0	0	0

Table A.2. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) under model 1. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Mixture Phenotypes Model 1			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.959	0.992	0.889	0.978	0.798	0.964	**1.408**	**1.120**	0.995	0.991	0.796	0.863
		0.0001	0.830	0.920	0.760	0.940	0.730	0.910	**1.500**	**1.240**	1.040	1.050	0.650	0.880
	4000	0.001	**1.087**	**1.094**	1.014	1.005	0.944	1.001	**1.282**	**1.142**	1.061	1.001	0.958	1.051
		0.0001	1.080	1.040	0.950	1.080	1.020	1.140	**1.450**	1.130	1.050	0.990	1.020	1.070
80	2000	0.001	0.926	0.970	0.839	0.931	0.801	0.957	**1.790**	**1.279**	0.995	0.972	0.775	0.842
		0.0001	0.760	0.850	0.900	1.020	0.680	0.990	**1.960**	**1.260**	0.930	0.930	0.640	0.750
	4000	0.001	1.061	1.031	0.960	0.952	0.884	0.975	**1.388**	**1.166**	0.985	0.936	0.894	0.980
		0.0001	1.090	1.040	0.930	1.120	0.790	1.090	**1.350**	1.160	1.040	0.910	0.810	1.020
100	2000	0.001	0.902	0.943	0.800	0.889	0.703	0.930	**2.147**	**1.291**	0.935	0.927	0.706	0.815
		0.0001	0.870	0.790	0.840	0.850	0.610	0.830	**2.440**	**1.500**	0.790	0.860	0.620	0.870
	4000	0.001	0.985	1.032	0.940	0.970	0.887	0.977	**1.523**	**1.155**	0.941	1.003	0.890	0.946
		0.0001	0.890	0.980	1.010	1.100	0.630	1.030	**1.390**	1.110	1.110	1.010	0.710	0.880

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O." represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.3. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) under model 2. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Mixture Phenotypes Model 2			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.968	0.988	0.899	0.936	0.851	0.956	**1.501**	**1.206**	1.006	0.974	0.848	0.909
		0.0001	0.910	0.840	0.870	0.870	0.780	1.060	**1.540**	**1.300**	1.080	0.940	0.790	0.810
	4000	0.001	1.047	1.057	0.976	0.969	0.930	0.993	**1.230**	**1.122**	0.984	0.972	0.915	0.977
		0.0001	0.960	1.010	1.030	1.000	1.000	1.000	**1.380**	1.180	1.110	1.000	0.970	1.080
80	2000	0.001	0.894	0.997	0.841	0.967	0.757	0.944	**1.738**	**1.253**	1.051	1.018	0.751	0.847
		0.0001	0.810	1.010	0.730	0.880	0.760	0.890	**1.990**	**1.580**	1.160	1.150	0.770	0.950
	4000	0.001	0.967	0.992	0.965	0.912	0.879	0.955	**1.300**	**1.095**	1.030	1.011	0.861	0.908
		0.0001	0.980	0.970	0.840	0.920	0.700	0.980	**1.270**	1.120	0.990	0.970	0.780	0.740
100	2000	0.001	0.858	0.939	0.821	0.885	0.707	0.908	**2.128**	**1.344**	1.015	0.930	0.704	0.782
		0.0001	0.810	0.960	0.730	0.880	0.570	1.090	**2.390**	**1.430**	0.980	0.850	0.610	0.790
	4000	0.001	0.995	0.984	0.886	0.962	0.881	0.980	**1.461**	**1.193**	1.046	0.957	0.870	0.922
		0.0001	1.080	0.990	0.970	0.890	0.930	0.990	**1.900**	1.070	1.030	1.040	0.910	0.760

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.4. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) under model 3. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Mixture Phenotypes Model 3			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.958	1.004	0.893	0.997	0.826	0.919	**1.481**	**1.155**	1.014	1.073	0.803	0.941
		0.0001	0.880	1.000	0.800	1.170	0.620	0.990	**1.530**	**1.230**	0.930	1.170	0.610	0.790
	4000	0.001	1.030	1.052	0.941	0.940	0.909	0.976	**1.210**	**1.114**	0.984	1.064	0.903	0.981
		0.0001	0.700	1.050	0.840	1.010	0.840	1.080	**1.420**	1.180	1.010	1.210	0.960	0.840
80	2000	0.001	0.949	0.952	0.843	0.909	0.769	0.882	**1.741**	**1.194**	1.009	**1.099**	0.739	0.903
		0.0001	0.710	0.850	0.790	0.760	0.580	0.740	**2.060**	**1.250**	0.990	1.070	0.580	0.740
	4000	0.001	1.006	0.965	0.903	0.912	0.881	0.907	**1.358**	**1.131**	0.915	1.000	0.887	0.951
		0.0001	0.890	0.920	0.900	1.070	0.810	0.910	**1.470**	1.070	0.840	1.090	0.850	1.020
100	2000	0.001	0.907	0.946	0.835	0.911	0.767	0.884	**2.171**	**1.383**	0.979	**1.071**	0.751	0.892
		0.0001	0.830	0.890	0.670	0.810	0.740	0.780	**2.760**	**1.430**	0.840	1.070	0.680	0.770
	4000	0.001	1.005	0.961	0.902	0.967	0.833	0.925	**1.393**	**1.170**	0.977	1.038	0.797	0.927
		0.0001	0.900	0.730	0.950	0.950	0.630	0.770	**1.410**	0.940	0.960	1.060	0.650	0.710

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.5. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) under model 4. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Mixture Phenotypes Model 4			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.935	0.983	0.872	0.982	0.783	0.923	**1.417**	**1.098**	0.988	**1.071**	0.789	0.907
		0.0001	0.890	0.980	0.930	0.800	0.650	0.950	**1.460**	**1.260**	0.980	**1.370**	0.690	0.850
	4000	0.001	1.019	0.979	0.915	0.959	0.923	0.943	**1.183**	**1.084**	0.975	1.060	0.870	0.935
		0.0001	1.120	1.090	0.870	0.890	0.940	1.020	**1.320**	1.190	0.940	0.990	0.860	1.000
80	2000	0.001	0.905	0.974	0.910	0.923	0.785	0.891	**1.778**	**1.250**	1.028	0.994	0.789	0.900
		0.0001	0.860	0.860	0.760	0.880	0.550	0.910	**1.870**	**1.290**	1.020	1.170	0.510	0.820
	4000	0.001	1.018	1.024	0.936	0.983	0.887	1.005	**1.338**	**1.116**	1.012	1.024	0.887	0.957
		0.0001	0.870	0.940	0.790	0.980	1.000	0.990	**1.700**	1.060	0.970	1.090	1.050	0.860
100	2000	0.001	0.902	0.996	0.851	0.909	0.739	0.918	**2.186**	**1.311**	0.981	1.022	0.750	0.858
		0.0001	1.000	0.890	0.830	1.010	0.680	0.810	**2.750**	**1.450**	0.840	1.020	0.770	0.680
	4000	0.001	1.000	1.043	0.902	0.957	0.855	0.969	**1.537**	**1.148**	0.989	1.027	0.876	0.933
		0.0001	0.980	1.040	1.000	1.030	0.970	1.070	**1.760**	1.180	0.980	1.020	1.000	1.030

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.6. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) under model 5. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Mixture Phenotypes Model 5			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	1.017	1.011	0.888	0.948	0.872	0.908	**1.451**	**1.154**	1.015	1.017	0.816	0.926
		0.0001	0.780	1.060	0.780	1.000	0.770	1.040	**1.580**	**1.240**	1.060	0.890	0.750	0.890
	4000	0.001	1.047	1.061	0.919	0.968	0.926	0.997	**1.225**	**1.069**	1.024	1.057	0.918	0.959
		0.0001	0.980	1.060	0.900	0.950	0.930	1.070	**1.350**	1.150	1.180	1.180	0.840	0.880
80	2000	0.001	0.973	1.006	0.852	0.929	0.854	0.959	**1.846**	**1.283**	1.028	1.056	0.837	0.891
		0.0001	0.970	0.840	0.930	1.050	0.750	0.940	**2.080**	**1.220**	1.110	**1.370**	0.640	0.720
	4000	0.001	1.032	0.998	0.893	0.956	0.873	0.911	**1.347**	**1.087**	0.977	1.033	0.872	0.928
		0.0001	1.100	0.950	0.930	0.970	0.860	0.850	**1.250**	1.070	1.090	1.120	0.850	0.930
100	2000	0.001	0.843	0.964	0.834	0.891	0.706	0.837	**2.103**	**1.266**	0.978	1.007	0.700	0.824
		0.0001	0.790	0.980	0.740	0.760	0.560	0.890	**2.350**	**1.360**	1.050	1.100	0.540	0.760
	4000	0.001	0.937	1.003	0.931	0.943	0.884	0.935	**1.483**	**1.100**	1.028	1.026	0.861	0.888
		0.0001	0.880	0.990	0.860	1.020	0.660	0.900	**1.590**	1.110	0.970	1.020	0.690	0.850

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.7. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) under model 6. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Mixture Phenotypes Model 6			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.949	0.980	0.840	0.958	0.868	0.891	**1.460**	**1.143**	1.002	**1.124**	0.804	0.898
		0.0001	0.960	1.020	0.830	1.000	0.830	0.940	**1.660**	**1.250**	0.900	1.030	0.770	0.850
	4000	0.001	1.041	1.036	0.953	0.996	0.913	0.911	**1.212**	1.025	0.985	**1.130**	0.882	0.952
		0.0001	0.990	0.980	0.950	1.120	0.780	0.870	**1.220**	1.100	1.030	**1.320**	0.820	0.940
80	2000	0.001	0.921	0.952	0.848	0.965	0.787	0.867	**1.761**	**1.193**	0.994	**1.103**	0.754	0.887
		0.0001	0.800	0.890	0.780	0.800	0.630	0.870	**1.840**	**1.310**	0.940	1.050	0.540	0.730
	4000	0.001	0.989	1.040	0.974	1.021	0.923	0.926	**1.363**	1.061	1.026	**1.126**	0.917	0.898
		0.0001	0.820	0.950	0.880	0.950	0.770	0.930	**1.420**	1.130	1.040	**1.360**	0.900	0.920
100	2000	0.001	0.885	0.954	0.815	0.961	0.714	0.895	**2.097**	**1.317**	0.947	1.048	0.656	0.898
		0.0001	0.780	1.060	0.710	0.940	0.700	1.020	**2.460**	**1.400**	0.960	**1.260**	0.640	0.880
	4000	0.001	0.974	0.976	0.915	0.924	0.881	0.897	**1.445**	**1.069**	0.994	**1.137**	0.844	0.861
		0.0001	0.980	1.160	0.810	0.960	0.920	0.900	**1.800**	1.040	0.910	**1.480**	0.940	0.970

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.8. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 binary phenotypes (with extremely unbalanced case-control ratios) under model 1. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Binary Phenotypes Model 1			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.965	0.973	0.657	0.715	0.792	0.811	**3.865**	**2.433**	0.960	1.014	0.550	0.531
		0.0001	1.020	1.130	0.510	0.680	0.830	0.820	**5.310**	**2.760**	1.190	**1.330**	0.400	0.440
	4000	0.001	1.027	1.011	0.748	0.846	0.871	0.852	**2.251**	**1.695**	0.988	1.013	0.766	0.747
		0.0001	1.030	1.120	0.690	0.740	0.860	0.810	**2.510**	**1.870**	1.010	1.070	0.600	0.610
80	2000	0.001	0.969	1.011	0.659	0.682	0.747	0.775	**5.638**	**2.972**	0.924	**1.051**	0.550	0.530
		0.0001	0.870	0.910	0.570	0.620	0.780	0.630	**8.280**	**3.430**	0.960	1.190	0.500	0.320
	4000	0.001	1.000	1.035	0.790	0.778	0.900	0.846	**2.820**	**1.964**	0.969	0.975	0.785	0.761
		0.0001	**1.240**	1.190	0.740	0.760	0.990	0.900	**3.620**	**2.290**	0.940	1.180	0.670	0.680
100	2000	0.001	0.965	1.053	0.663	0.702	0.819	0.826	**8.393**	**3.867**	0.926	1.016	0.553	0.580
		0.0001	1.000	1.110	0.600	0.710	0.740	0.910	**13.66**	**5.150**	1.020	1.190	0.470	0.500
	4000	0.001	1.034	1.061	0.730	0.790	0.847	0.866	**3.454**	**2.212**	0.971	1.014	0.681	0.728
		0.0001	1.070	1.120	0.640	0.840	0.810	0.880	**4.500**	**2.670**	**1.030**	**1.220**	0.600	0.490

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.9. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 binary phenotypes (with extremely unbalanced case-control ratios) under model 2. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Binary Phenotypes Model 2			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.951	1.034	0.729	0.712	0.784	0.776	**4.020**	**2.401**	0.910	**1.070**	0.585	0.549
		0.0001	0.900	1.100	0.590	0.730	0.670	0.830	**5.190**	**2.630**	0.800	**1.420**	0.420	0.440
	4000	0.001	0.994	1.062	0.835	0.852	0.887	0.878	**2.181**	**1.703**	0.983	1.006	0.730	0.742
		0.0001	1.170	1.100	0.720	0.950	0.910	0.910	**2.640**	**1.890**	0.920	1.130	0.520	0.630
80	2000	0.001	0.977	1.061	0.671	0.715	0.807	0.793	**5.751**	**2.996**	0.897	**1.125**	0.543	0.542
		0.0001	1.130	1.170	0.570	0.690	0.870	0.880	**8.540**	**3.650**	1.050	**1.440**	0.430	0.370
	4000	0.001	1.027	1.054	0.811	0.829	0.904	0.823	**2.722**	**1.945**	0.927	1.037	0.717	0.761
		0.0001	1.170	1.090	0.610	0.940	0.920	0.850	**3.440**	**2.330**	1.090	**1.300**	0.750	0.520
100	2000	0.001	0.992	1.029	0.675	0.695	0.791	0.789	**8.261**	**3.754**	0.906	1.036	0.506	0.543
		0.0001	1.000	1.180	0.600	0.630	0.810	0.930	**12.77**	**4.530**	0.930	**1.370**	0.490	0.490
	4000	0.001	1.061	1.090	0.814	0.817	0.915	0.874	**3.458**	**2.279**	0.975	1.036	0.743	0.764
		0.0001	1.120	1.100	0.860	0.750	0.980	0.970	**4.380**	**2.730**	1.040	1.170	0.610	0.750

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.10. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 binary phenotypes (with extremely unbalanced case-control ratios) under model 3. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Binary Phenotypes Model 3			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.949	1.003	0.704	0.730	0.789	0.761	**3.525**	**2.089**	0.870	1.054	0.571	0.578
		0.0001	1.040	1.160	0.720	0.820	0.850	0.910	**4.610**	**2.390**	0.870	1.120	0.500	0.680
	4000	0.001	0.970	1.040	0.786	0.905	0.935	0.903	**2.076**	**1.494**	0.969	1.014	0.786	0.727
		0.0001	1.050	1.100	0.760	0.800	0.950	0.790	**2.670**	**1.580**	0.890	**1.210**	0.770	0.710
80	2000	0.001	0.940	1.044	0.632	0.689	0.773	0.855	**4.629**	**2.299**	0.855	1.056	0.519	0.544
		0.0001	1.070	1.120	0.660	0.840	0.700	1.050	**6.390**	**2.770**	0.840	1.190	0.430	0.460
	4000	0.001	1.009	1.061	0.770	0.850	0.925	0.899	**2.358**	**1.617**	0.954	**1.093**	0.728	0.715
		0.0001	0.980	1.120	0.780	0.900	0.930	0.820	**2.900**	**1.690**	1.000	1.160	0.740	0.580
100	2000	0.001	0.968	1.050	0.680	0.701	0.731	0.811	**6.697**	**3.000**	0.876	1.058	0.533	0.570
		0.0001	0.930	1.110	0.720	0.710	0.680	0.920	**10.29**	**3.760**	0.790	1.150	0.400	0.430
	4000	0.001	0.998	1.060	0.764	0.794	0.898	0.873	**2.978**	**1.831**	0.987	**1.093**	0.690	0.715
		0.0001	1.140	1.190	0.650	0.820	0.750	0.970	**3.330**	**2.040**	0.930	1.190	0.600	0.760

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.11. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 binary phenotypes (with extremely unbalanced case-control ratios) under model 4. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Binary Phenotypes Model 4			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.978	1.036	0.656	0.750	0.801	0.892	**3.362**	**1.986**	0.924	1.055	0.537	0.574
		0.0001	1.110	1.120	0.700	0.710	0.720	0.830	**4.040**	**2.430**	1.050	1.440	0.340	0.470
	4000	0.001	1.031	1.049	0.827	0.824	0.929	0.853	**2.068**	**1.527**	0.968	1.061	0.724	0.759
		0.0001	1.080	1.100	0.800	0.790	1.050	0.800	**2.390**	**1.670**	1.180	1.170	0.650	0.600
80	2000	0.001	0.946	1.059	0.666	0.736	0.792	0.853	**4.719**	**2.412**	0.896	**1.119**	0.543	0.585
		0.0001	1.050	1.110	0.620	0.810	0.860	0.850	**6.410**	**2.780**	0.830	**1.420**	0.420	0.500
	4000	0.001	0.944	1.061	0.786	0.842	0.895	0.913	**2.417**	**1.707**	0.919	1.053	0.750	0.769
		0.0001	1.070	**1.250**	0.630	0.860	0.830	0.880	**3.170**	**1.930**	0.960	1.120	0.740	0.800
100	2000	0.001	1.023	1.037	0.731	0.748	0.823	0.847	**6.615**	**2.861**	0.944	1.074	0.526	0.542
		0.0001	1.090	1.170	0.740	0.780	0.730	1.040	**10.04**	**3.580**	0.950	**1.230**	0.430	0.470
	4000	0.001	1.010	1.057	0.761	0.802	0.905	0.914	**2.986**	**1.760**	0.909	1.053	0.720	0.729
		0.0001	0.960	1.120	0.660	0.590	0.790	0.930	**3.770**	**1.930**	0.900	**1.290**	0.680	0.650

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.12. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 binary phenotypes (with extremely unbalanced case-control ratios) under model 5. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Binary Phenotypes Model 5			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.974	1.044	0.727	0.715	0.866	0.869	**3.530**	**2.140**	0.973	1.058	0.560	0.541
		0.0001	1.020	0.960	0.630	0.620	0.830	0.780	**4.460**	**2.350**	0.930	1.180	0.330	0.380
	4000	0.001	1.025	1.056	0.804	0.837	0.883	0.899	**2.053**	**1.540**	0.944	1.062	0.734	0.746
		0.0001	1.160	1.140	0.670	0.760	0.980	0.860	**2.570**	**1.510**	1.030	1.180	0.730	0.470
80	2000	0.001	0.965	1.057	0.687	0.740	0.807	0.875	**4.909**	**2.588**	0.909	**1.130**	0.522	0.577
		0.0001	1.040	1.110	0.570	0.680	0.870	0.950	**6.400**	**3.210**	0.820	**1.250**	0.410	0.450
	4000	0.001	0.993	1.036	0.838	0.819	0.879	0.882	**2.450**	**1.732**	0.924	1.056	0.713	0.722
		0.0001	1.120	1.150	0.610	0.930	1.070	0.990	**3.130**	**2.070**	1.000	**1.350**	0.600	0.680
100	2000	0.001	0.871	1.060	0.714	0.736	0.746	0.807	**6.884**	**3.209**	0.874	1.058	0.550	0.537
		0.0001	1.040	1.180	0.670	0.690	0.600	0.990	**10.67**	**3.870**	0.880	1.130	0.360	0.460
	4000	0.001	0.969	1.099	0.819	0.804	0.915	0.897	**3.179**	**1.904**	0.914	1.050	0.786	0.733
		0.0001	1.040	1.160	0.710	0.970	0.770	1.030	**4.410**	**2.220**	1.070	1.160	0.550	0.600

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.13. The estimated type I error rates of the six multiple phenotype association tests divided by the nominal significance level for 60, 80, and 100 binary phenotypes (with extremely unbalanced case-control ratios) under model 6. The type I error rates are evaluated using 500 MC runs (equivalent to 10^6 replicates).

Binary Phenotypes Model 6			**ceCLC**		**CLC**		**HCLC**		**MultiPhen**		**O'Brien**		**Omnibus**	
K	Sample	α-level	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET	N.O.	NET
60	2000	0.001	0.995	1.061	0.719	0.852	0.838	0.900	**3.288**	**1.771**	0.907	1.024	0.566	0.537
		0.0001	1.130	1.180	0.710	0.780	0.800	1.010	**4.290**	**2.030**	1.000	1.140	0.410	0.470
	4000	0.001	1.004	1.006	0.849	0.943	0.890	0.964	**1.949**	**1.452**	0.967	**1.082**	0.726	0.739
		0.0001	1.120	1.130	0.770	1.020	1.070	1.120	**2.200**	**1.580**	1.070	**1.280**	0.710	0.850
80	2000	0.001	0.940	1.072	0.708	0.785	0.776	0.876	**4.481**	**2.167**	0.953	1.061	0.570	0.549
		0.0001	0.980	1.070	0.690	0.810	0.740	0.880	**6.110**	**2.520**	0.840	1.130	0.420	0.400
	4000	0.001	1.011	**1.129**	0.774	0.838	0.913	0.923	**2.303**	**1.549**	0.988	**1.150**	0.718	0.743
		0.0001	0.910	**1.240**	0.670	1.120	0.810	0.990	**3.130**	**1.910**	0.900	**1.360**	0.770	0.770
100	2000	0.001	0.938	1.081	0.684	0.744	0.776	0.892	**6.011**	**2.547**	0.927	1.012	0.499	0.534
		0.0001	0.890	1.160	0.600	0.810	0.590	1.040	**8.730**	**3.130**	0.940	1.120	0.390	0.410
	4000	0.001	0.974	1.026	0.783	0.862	0.889	0.951	**2.780**	**1.584**	0.997	1.033	0.721	0.723
		0.0001	0.890	1.180	0.680	1.010	0.900	1.080	**3.560**	**1.990**	0.910	1.110	0.680	0.660

Notes: bold-faced values indicate that the values are beyond the upper bounds of the 95% CIs. 95% CIs for type I error rates divided by nominal significance levels 0.001 and 0.0001 are (0.938, 1.062) and (0.804, 1.196), respectively. "N.O" represents the type I error rates calculated by the formula in Comparison 1 (Apply methods without considering network modules.). "NET" presents the type I error rates evaluated by the formula in Comparison 2 (Apply methods by considering network modules).

Table A.14. 33 unique SNPs identified by ceCLC for testing the association in NET (one SNP) or in N.O. (32 SNPs).

SNP	Position	Mapped gene	P value	Reported diseases	Reference
rs4148866*	chr12: 123425575	*ABCB9*	2.97E-08	-	-
rs13107325	chr4:102267552	*SLC39A8*	4.60E-10	M19.9 / M25.5 / M75.1	[10-12]
rs34333163	chr4:102361960	*SLC39A8*	2.84E-08	M19.9 / M25.5 / M75.1	[10-12]
rs9468413	chr6:28721895	-	2.91E-08	-	-
rs880638	chr6:28739135	-	4.02E-08	-	-
rs9257802	chr6:29375578	*OR5V1*	1.50E-08	-	-
rs1264362	chr6:30808813	*HCG20*	1.54E-09	M07.3	[13]
rs915664	chr6:30826840	*LINC00243*	1.02E-08	-	-
rs1264344	chr6:30832800	*LINC00243*	9.31E-09	-	-
rs1632854	chr6:31007872	*MUC21 / MUC22*	2.77E-08	M07.3 / M32.9 / M85.8	[13-15]
rs4713422	chr6:31032125	*MUC22*	8.63E-14	M07.3 / M85.8	[13,15]
rs10947121	chr6:31032220	*MUC22*	1.08E-13	M07.3 / M85.8	[13,15]
rs2233967	chr6:31113051	*C6orf15 / PSORS1C1*	4.61E-09	M31.4 / M34 / M35.2	[16-18]
rs1265086	chr6:31142105	*PSORS1C1 / PSORS1C2*	5.07E-12	M31.4 / M34 / M35.2	[16-18]
rs130071	chr6:31148433	*PSORS1C1 / POU5F1*	2.46E-10	M07.3 / M31.4 / M34 / M35.2	[13,16-18]
rs4516988	chr6:31208825	*HCG27*	7.30E-10	M32.9	[19]
rs4351302	chr6:31209144	*HCG27*	4.08E-10	M32.9	[19]
rs9295967	chr6:31216243	*HCG27*	6.65E-10	M32.9	[19]
rs9264733	chr6:31276437	*HLA-C / LINC02571*	1.10E-08	M07.3 / M31.4	[13,17]
rs3094682	chr6:31296684	*LINC02571*	2.58E-09	M07.3	[13]
rs2596472	chr6:31461190	*HCP5 / MICB*	1.33E-10	M33.2 / M60 / M62.9	[20-22]
rs3130615	chr6:31507636	*MICB*	1.79E-11	M60	[21]
rs3132468	chr6:31507709	*MICB*	1.57E-11	M60	[21]
rs3131635	chr6:31508357	*MICB*	1.13E-11	M60	[21]
rs1065076	chr6:31509904	*MICB*	1.26E-11	M60	[21]
rs2395045	chr6:31516740	*MICB*	1.07E-09	M60	[21]
rs3093999	chr6:31516773	*MICB*	8.32E-10	M60	[21]
rs3131631	chr6:31516906	*MICB*	9.63E-10	M60	[21]
rs2734574	chr6:31526111	*MICB*	5.30E-09	M60	[21]
rs6916921	chr6:31552649	*ATP6V1G2 / DDX39B / LTA*	2.28E-11	M30.3 / M60	[21,23]
rs915895	chr6:32222440	*NOTCH4*	1.94E-09	M06.9 / M07.3 / M31.4 / M32.9 / M34 / M62.9	[13,17,19,22,24]
rs915894	chr6:32222613	*NOTCH4*	2.14E-08	M06.9 / M07.3 / M31.4 / M32.9 / M34 / M62.9	[13,17,19,22,24]
rs443198	chr6:32222629	*NOTCH4*	1.73E-11	M06.9 / M07.3 / M31.4 / M32.9 / M34 / M62.9	[13,17,19,22,24]

Notes: "*" indicates the unique SNP identified by ceCLC in NET. Bold-faced SNPs are the lead SNPs in the colocalization analysis. Mapped gene denotes the gene that includes the corresponding SNP with a 20kb window region. P-value is calculated by ceCLC. The corresponding diseases with ICD-10 codes in reported diseases are listed in Table A.15.

Table A.15. ICD-10 codes and names of the14 reported diseases shown in Table A.14.

ICD-10	Disease	ICD-10	Disease
M06.9	rheumatoid arthritis	M33.2	Polymyositis
M07.3	psoriatic arthritis	M34	systemic sclerosis
M19.9	osteoarthritis	M35.2	Behcet's disease
M25.5	multisite chronic pain	M60	myositis
M30.3	Kawasaki disease	M62.9	appendicular lean mass
M31.4	Takayasu arteritis	M75.1	rotator cuff syndrome
M32.9	systemic lupus erythematosus	M85.8	disorders of bone density and structure

A.3 Supplementary Figures

Figure A.1. Flow chart of UK Biobank data preprocessing. *Pre-process on phenotype:* i. Select White British subjects (White British); ii. Remove individuals who are marked as outliers for heterozygosity or missing rates (Low Heterozygosity); iii. Exclude individuals who have been identified to have ten or more third-degree relatives or closer (Not Three-degree Relatives); iv. Remove individuals having very similar ancestry based on a principal component analysis of the genotypes (Similar Ancestry); v. Remove individuals based on removal by the UK Biobank (Removal by the UK Biobank). *Quality controls (QCs) on genotype:* Filter out genetic variants, with i. Missing rate larger than 5% ("--mind 0.05"), ii. Hardy-Weinberg equilibrium exact test p-values less than 10^{-6} ("--hwe 1e-6"), iii. Minor allele frequency (MAF) less than 5% ("--maf 0.05"). We also filter out individuals, with iv. Missing rate larger than 5% ("--mind 0.05") v. Individuals without sex ("--no-sex").

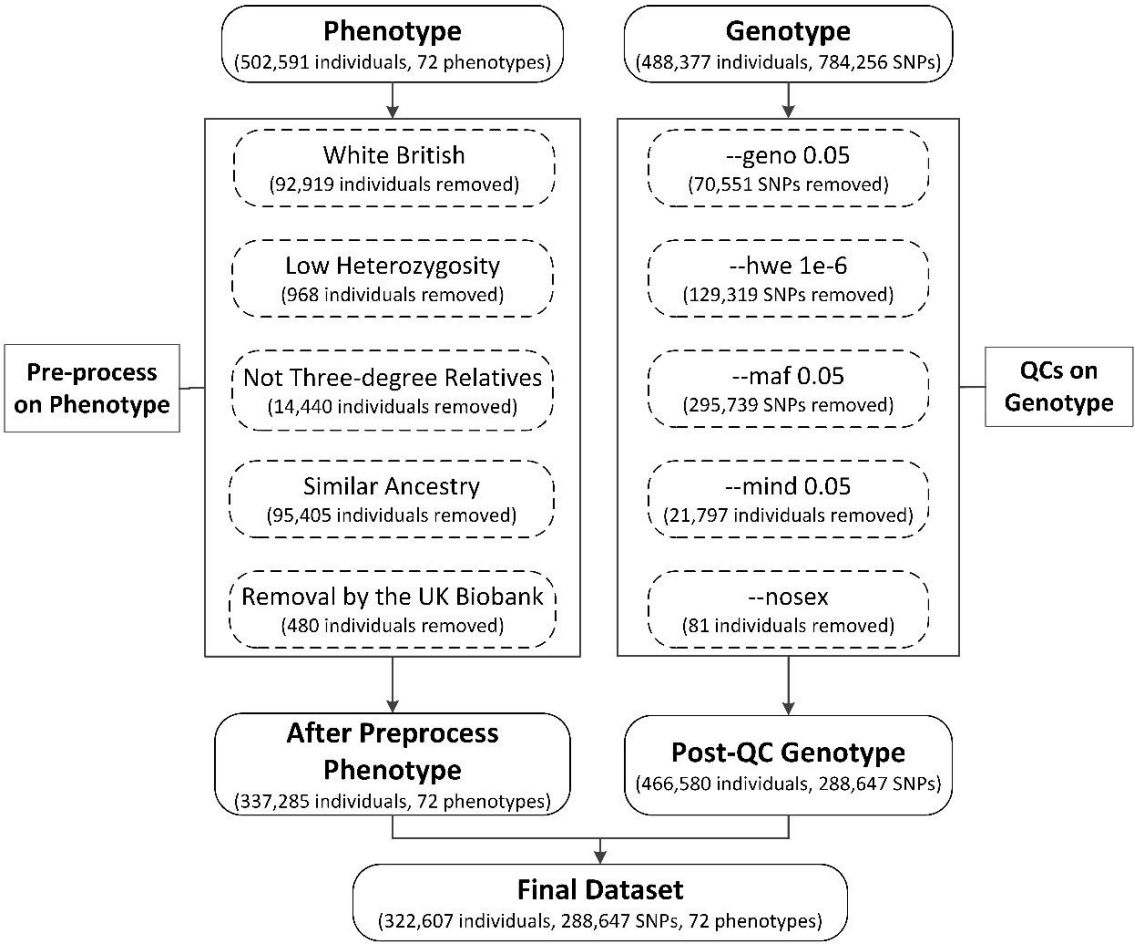


Figure A.2. Power comparisons of the six tests as a function of effect size β under six models. The number of mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) is 80 and the sample size is 2,000. The power of all of the six tests is evaluated using 10 MC runs.

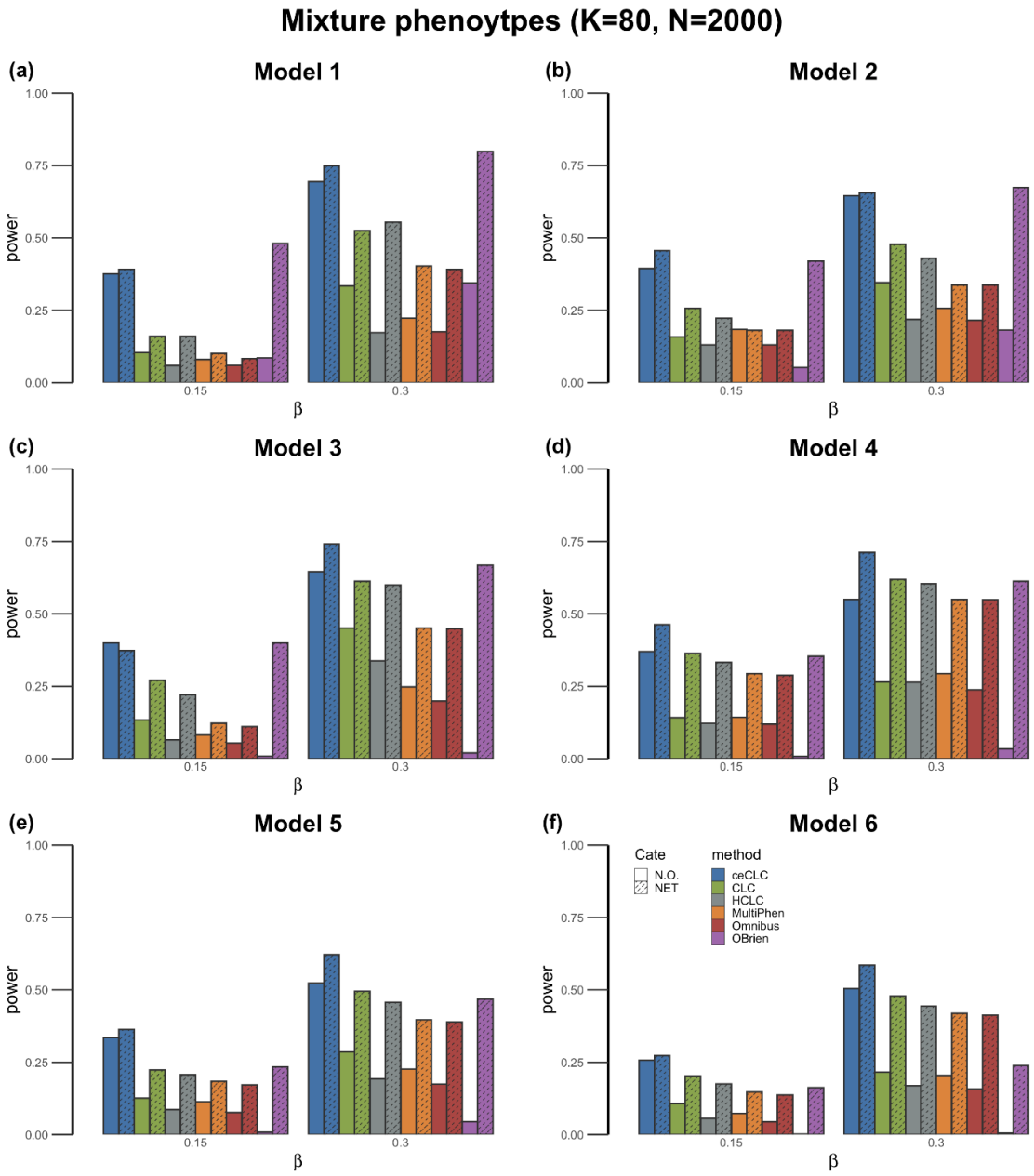


Figure A.3. Power comparisons of the six tests as a function of effect size β under six models. The number of mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) is 60 and the sample size is 2,000. The power of all of the six tests is evaluated using 10 MC runs.

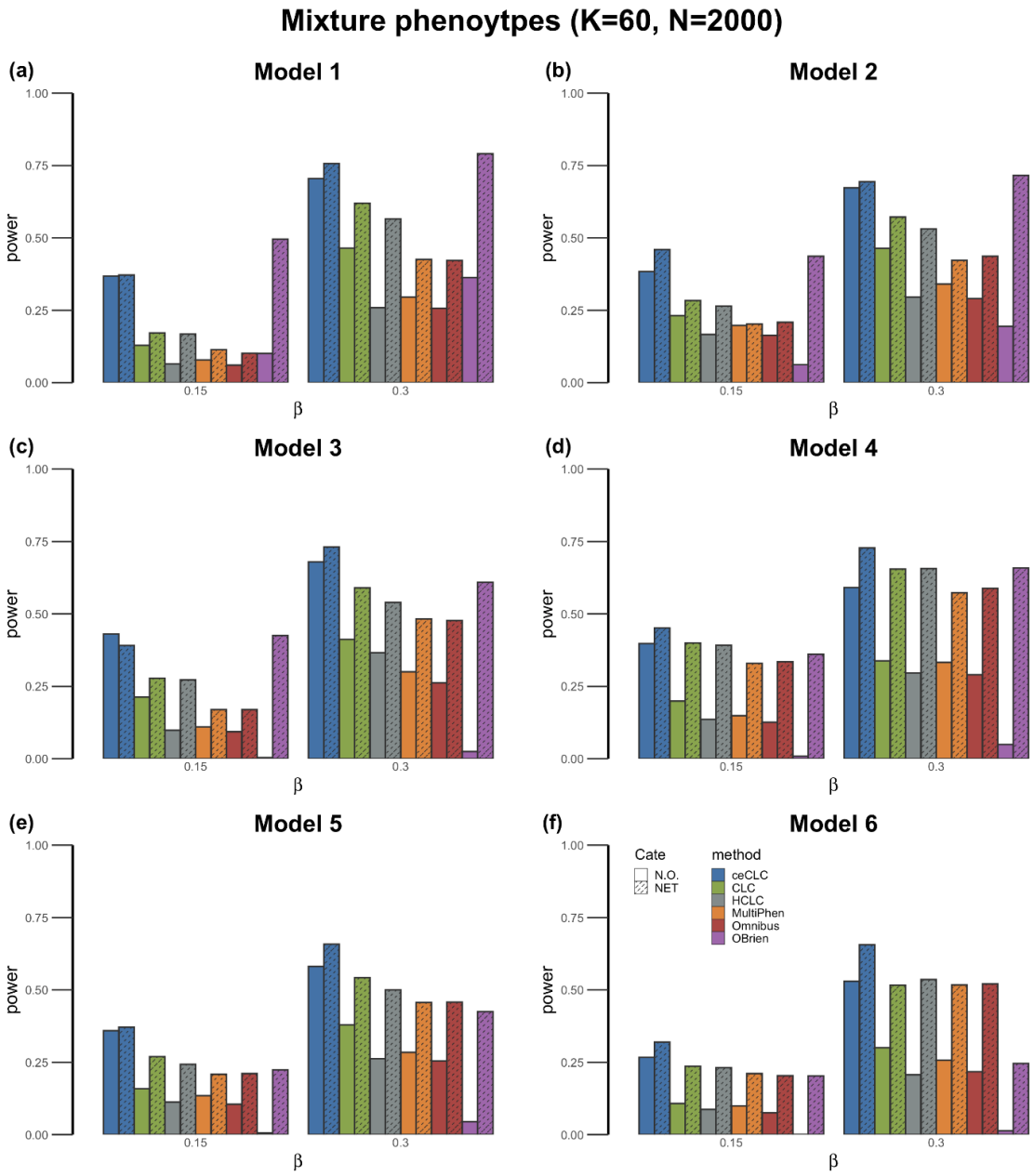


Figure A.4. Power comparisons of the six tests as a function of effect size β under six models. The number of mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) is 60 and the sample size is 4,000. The power of all of the six tests is evaluated using 10 MC runs.

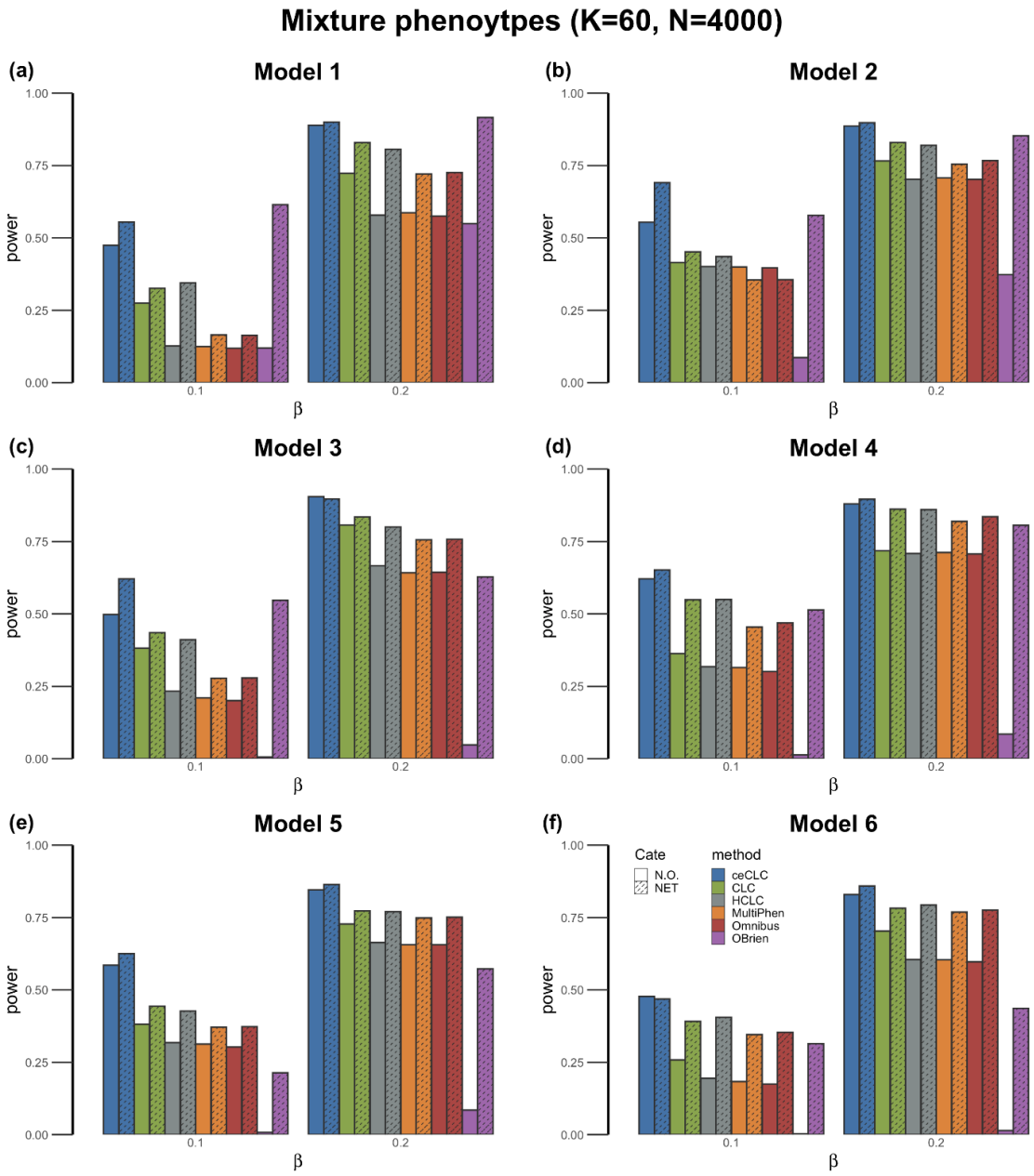


Figure A.5. Power comparisons of the six tests as a function of effect size β under six models. The number of mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) is 100 and the sample size is 2,000. The power of all of the six tests is evaluated using 10 MC runs.

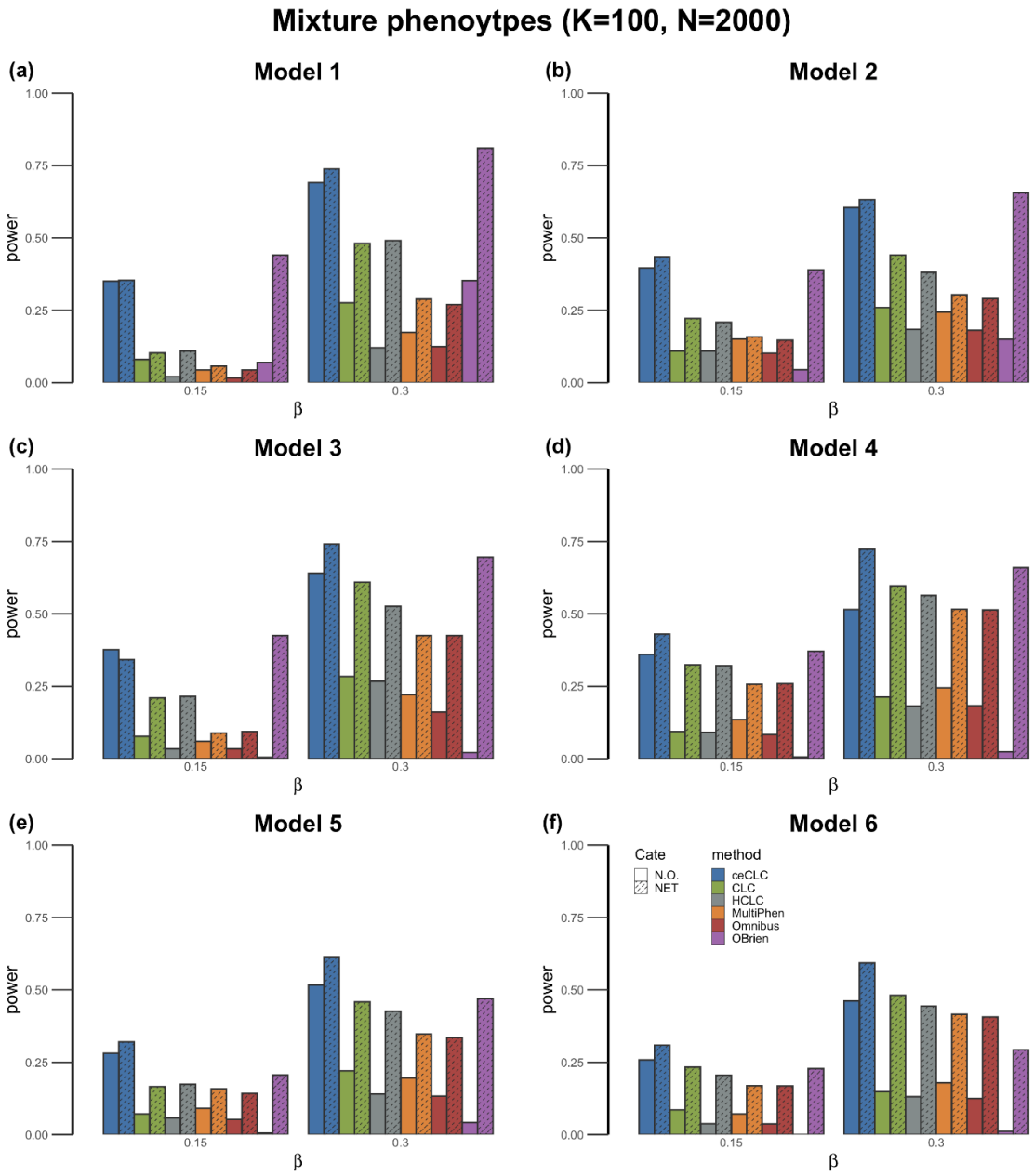


Figure A.6. Power comparisons of the six tests as a function of effect size β under six models. The number of mixture phenotypes (half continuous phenotypes and half binary phenotypes with balanced case-control ratios) is 100 and the sample size is 4,000. The power of all of the six tests is evaluated using 10 MC runs.

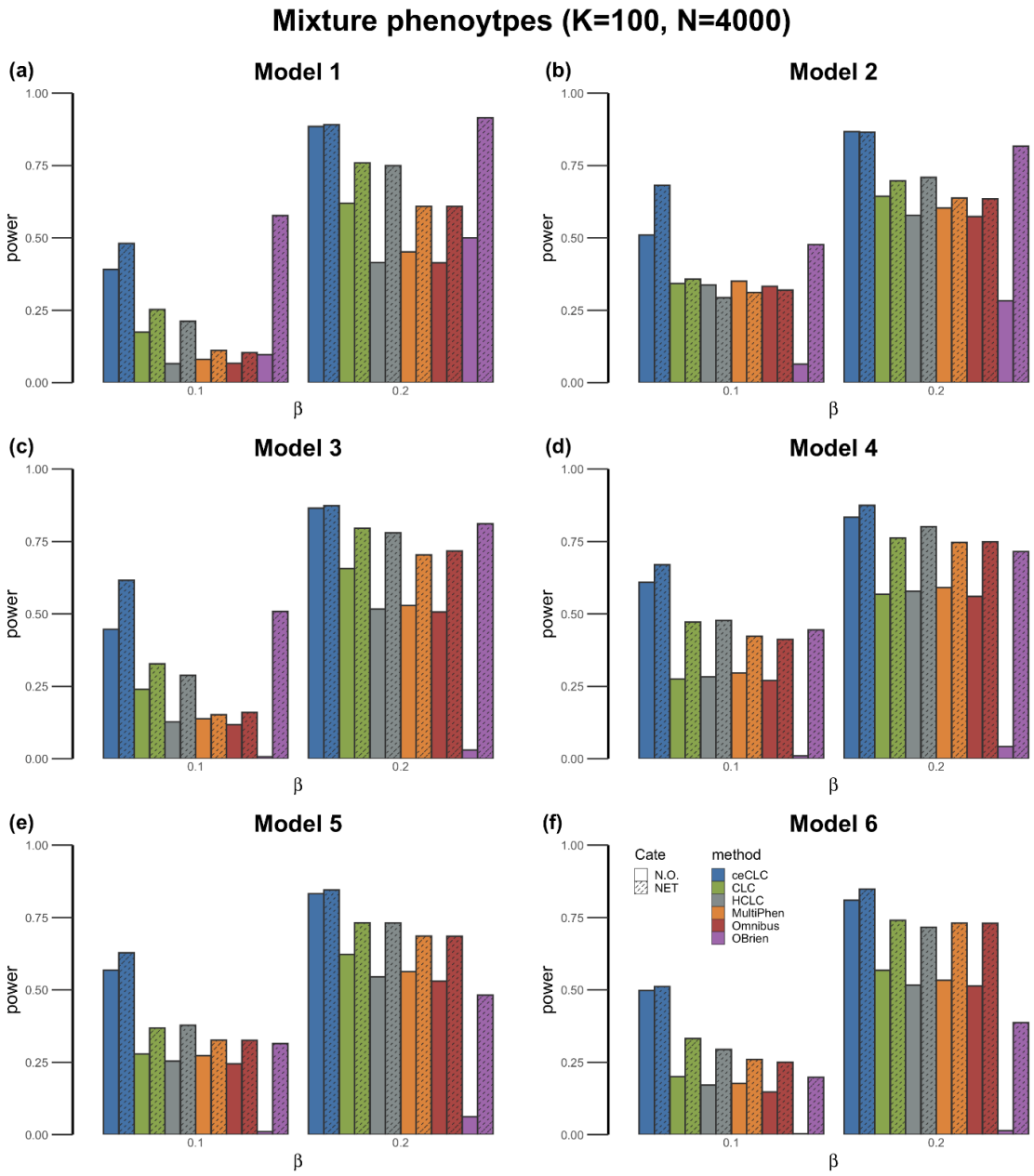


Figure A.7. Power comparisons of the six tests as a function of effect size β under the six models. The number of binary phenotypes (with extremely unbalanced case-control ratios) is 80 and the sample size is 20,000. The power of all of the six tests is evaluated using 10 MC runs.

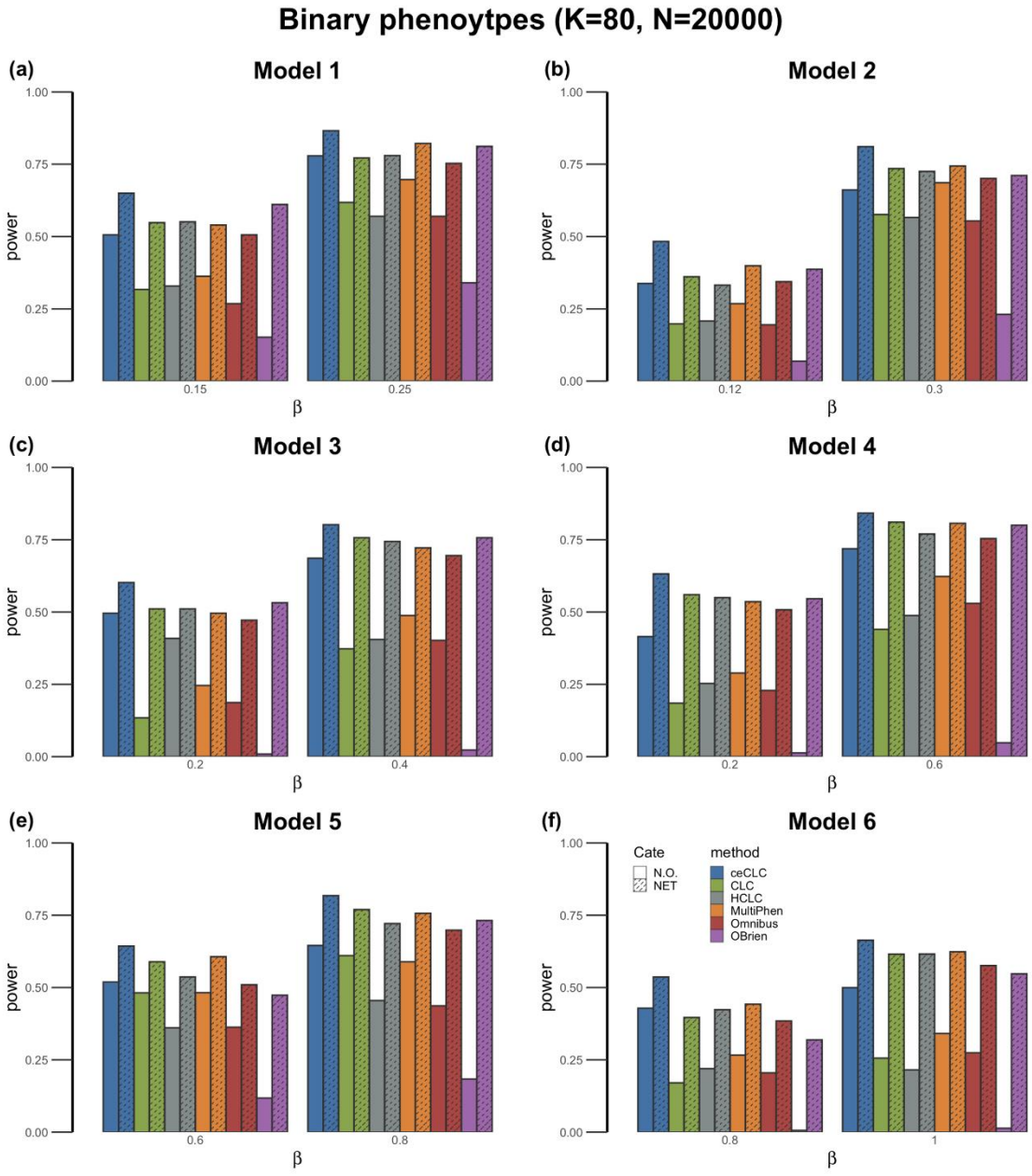


Figure A.8. Power comparisons of the six tests as a function of effect size β under six models. The number of binary phenotypes (with extremely unbalanced case-control ratios) is 80 and the sample size is 10,000. The power of all of the six tests is evaluated using 10 MC runs.

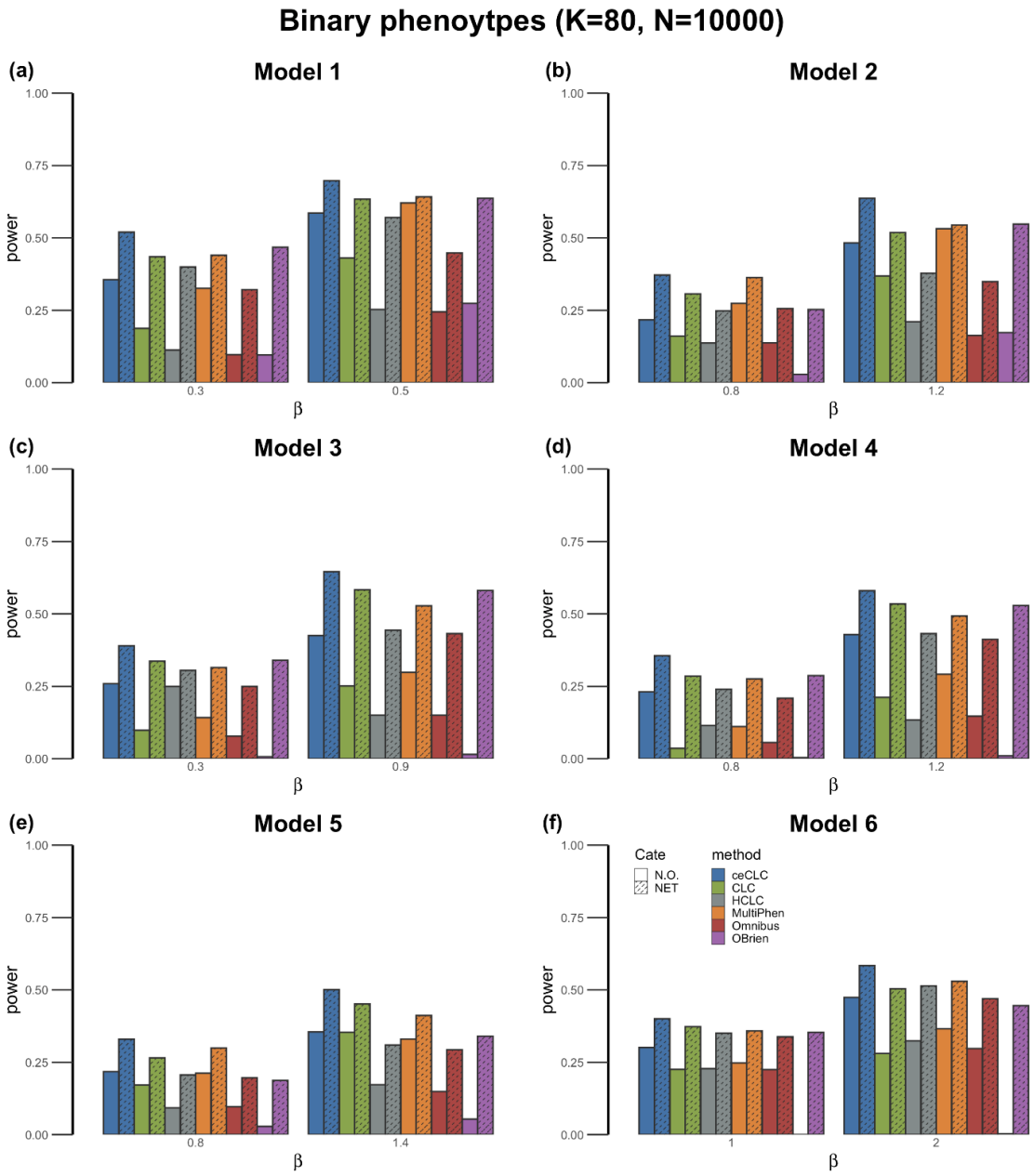


Figure A.9. Power comparisons of the six tests as a function of effect size β under six models. The number of binary phenotypes (with extremely unbalanced case-control ratios) is 60 and the sample size is 10,000. The power of all of the six tests is evaluated using 10 MC runs.

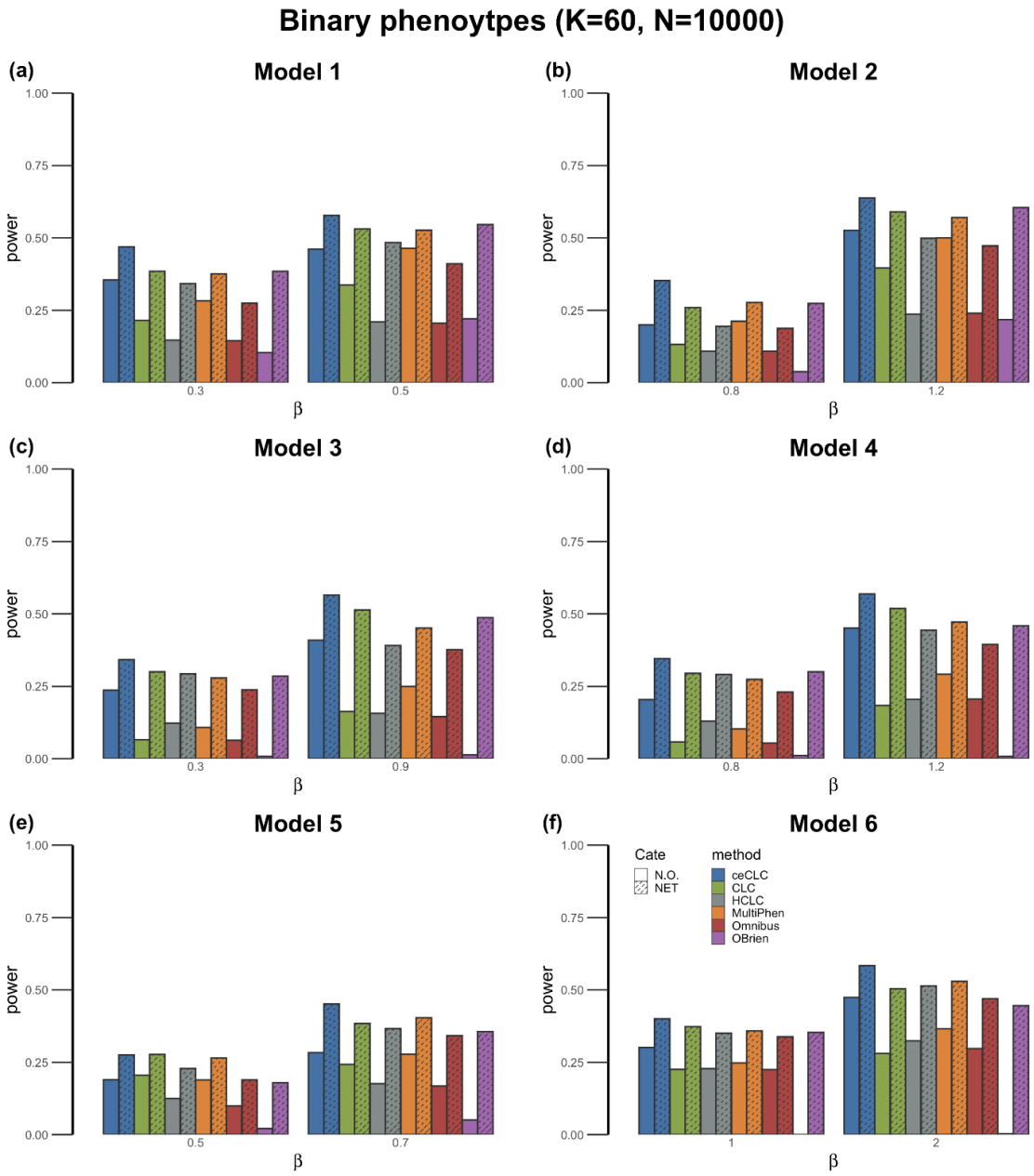


Figure A.10. Power comparisons of the six tests as a function of effect size β under six models. The number of binary phenotypes (with extremely unbalanced case-control ratios) is 60 and the sample size is 20,000. The power of all of the six tests is evaluated using 10 MC runs.

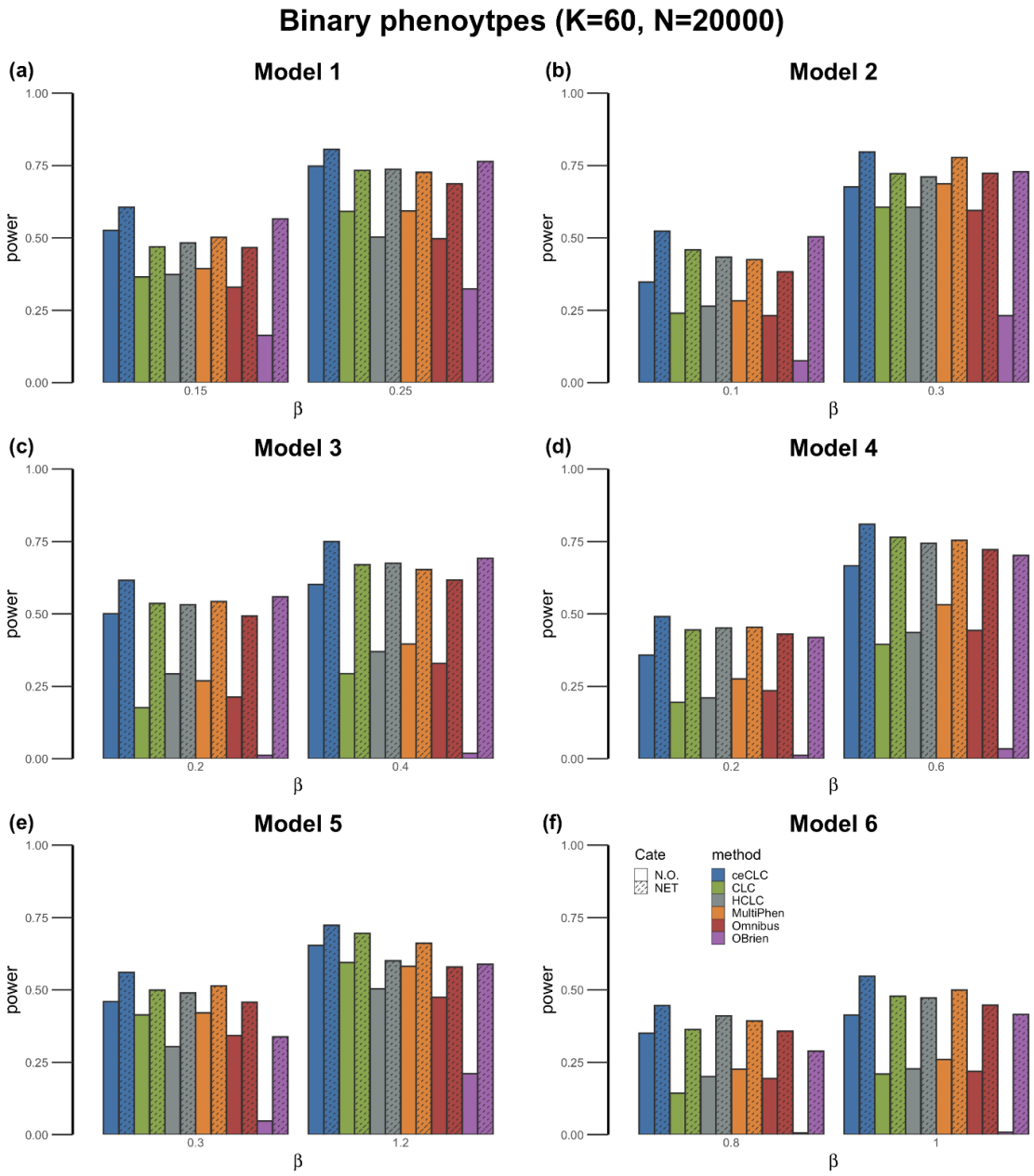


Figure A.11. Power comparisons of the six tests as a function of effect size β under six models. The number of binary phenotypes (with extremely unbalanced case-control ratios) is 100 and the sample size is 10,000. The power of all of the six tests is evaluated using 10 MC runs.

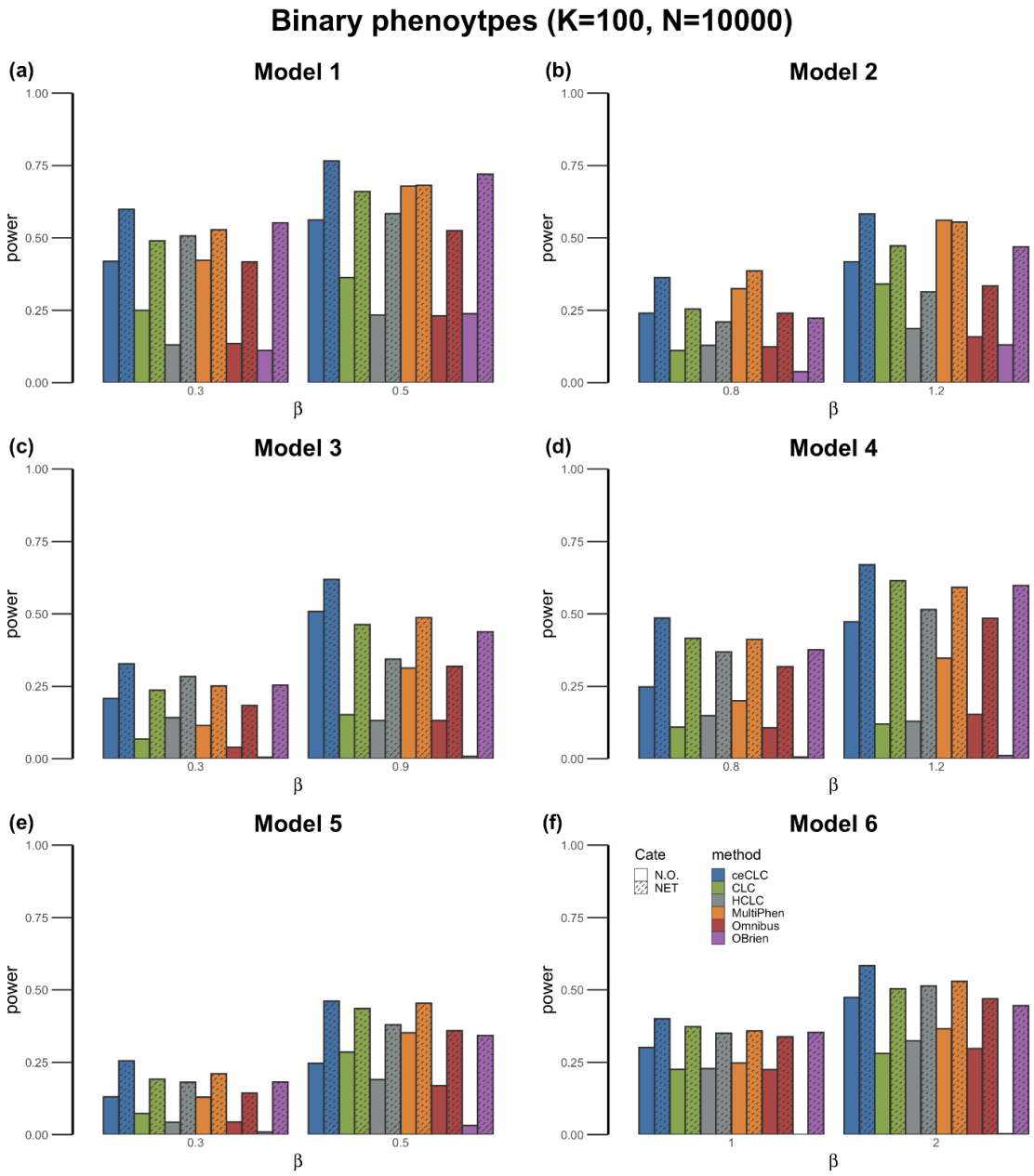


Figure A.12. Power comparisons of the six tests as a function of effect size β under six models. The number of binary phenotypes (with extremely unbalanced case-control ratios) 100 and the sample size is 20,000. The power of all of the six tests is evaluated using 10 MC runs.

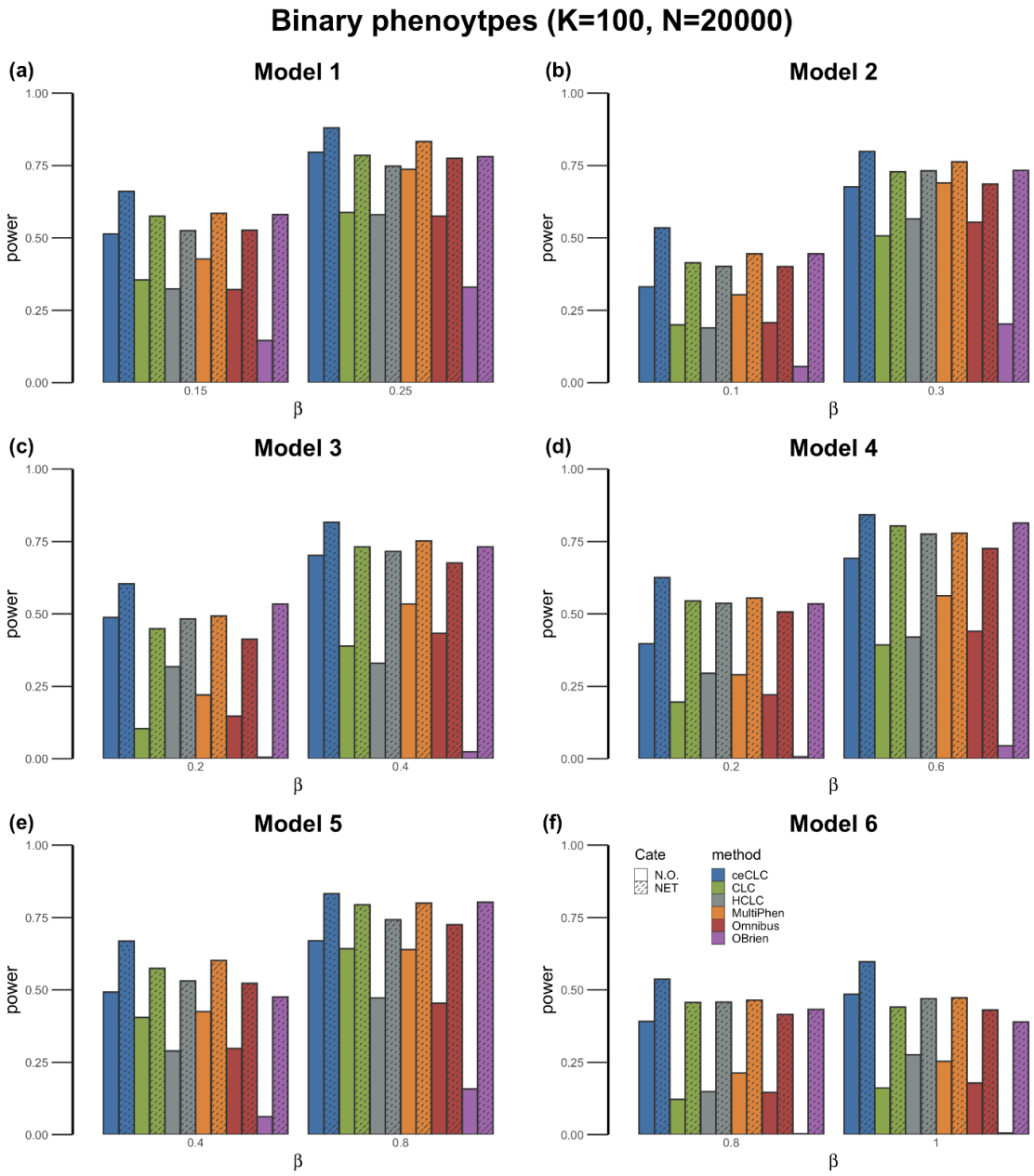


Figure A.13. The Manhattan plots of four different diseases based on the saddlepoint approximation. Systemic lupus erythematosus (M32.9), Sicca syndrome (M35.0), and Trigger finger (M65.3) are detected in Module III by our proposed GPN. Both Trigger finger (M65.3) and Synovitis and tenosynovitis (M65.9) are classified into the same ICD-codes category (M65). The horizontal red dashed line represents the threshold for commonly used genome-wide significance level 5×10^{-8}.

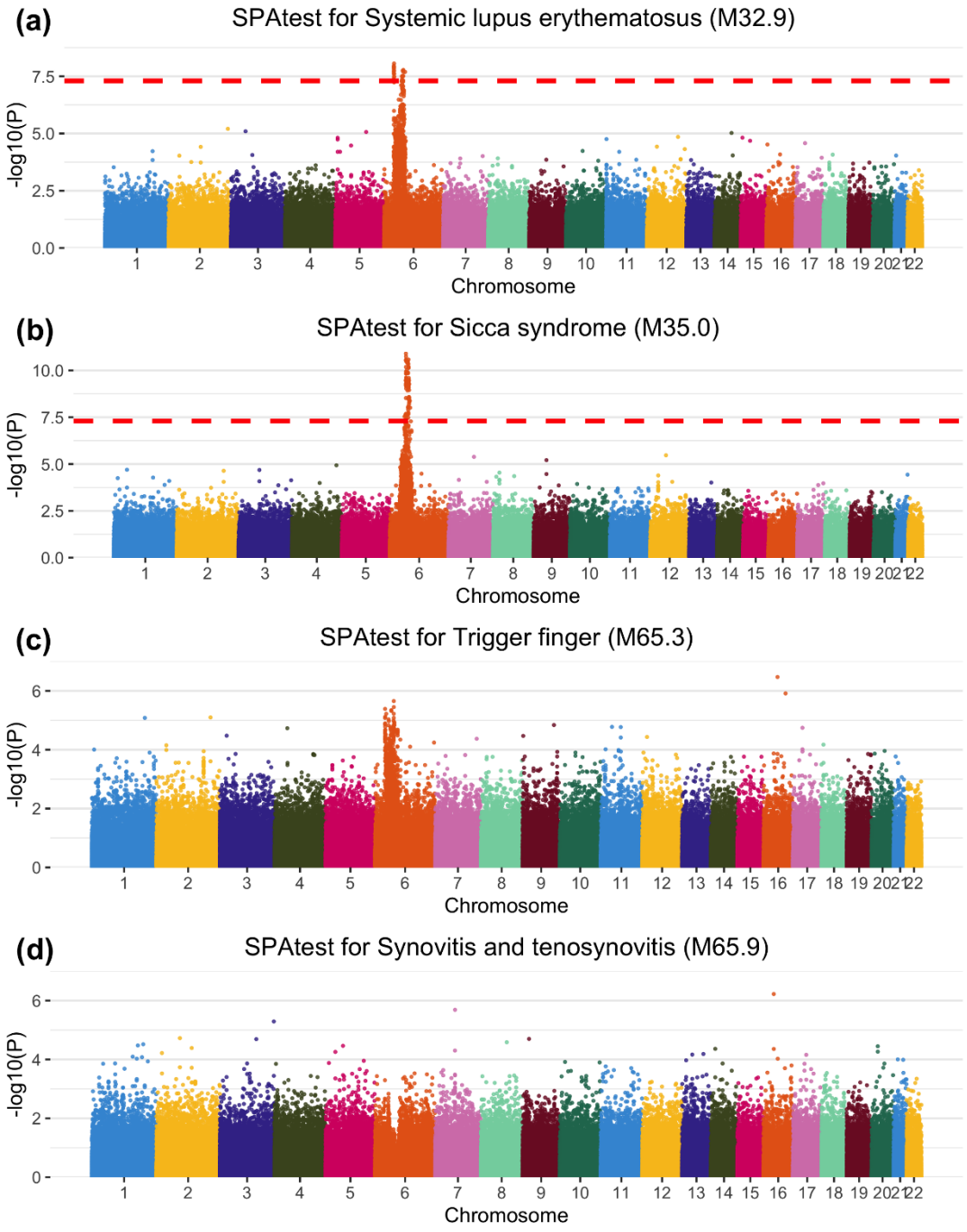


Figure A.14. Dendrogram of hierarchical clustering method based on the genetic correlation of phenotypes obtained by GPN and the phenotypic correlation estimated by LDSC, respectively.

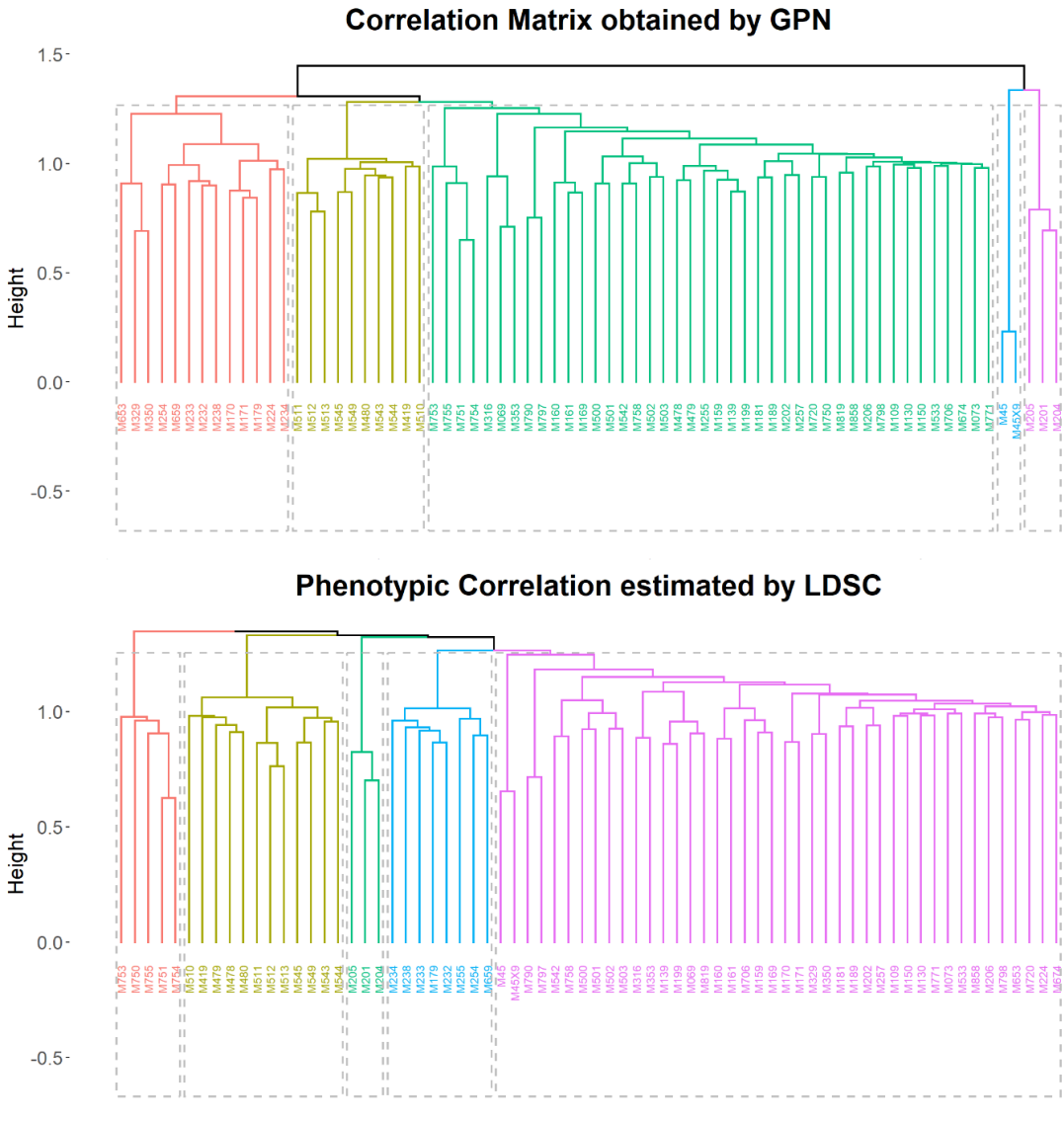


Figure A.15. Dendrogram of hierarchical clustering method based on the genetic correlation of phenotypes obtained by GPN and the genetic correlation estimated by LDSC, respectively.

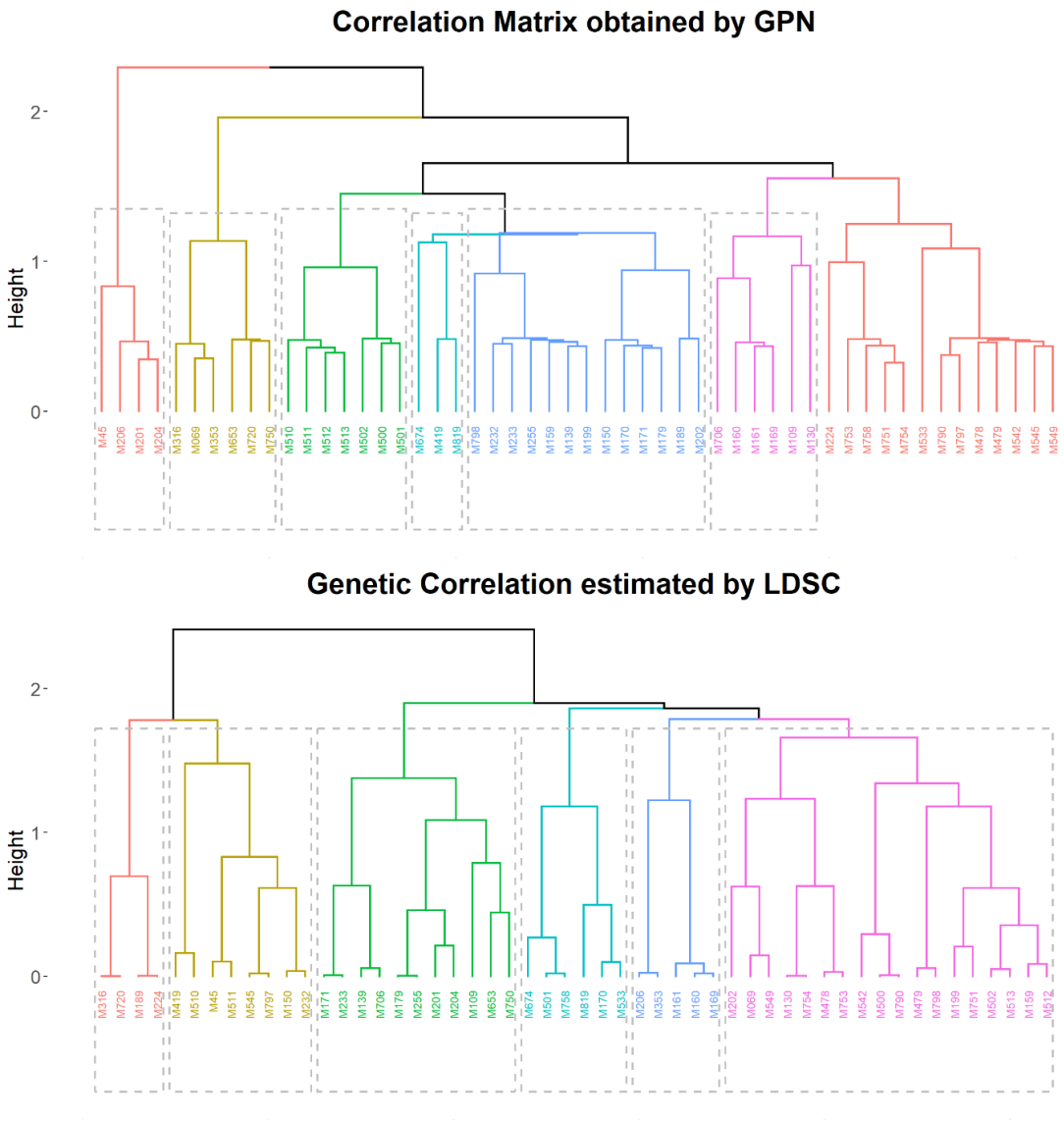


Figure A.16. The Venn diagrams of the number of significant SNPs identified by ceCLC, CLC, HCLC, O'Brien, and Omnibus in N.O. and NET.

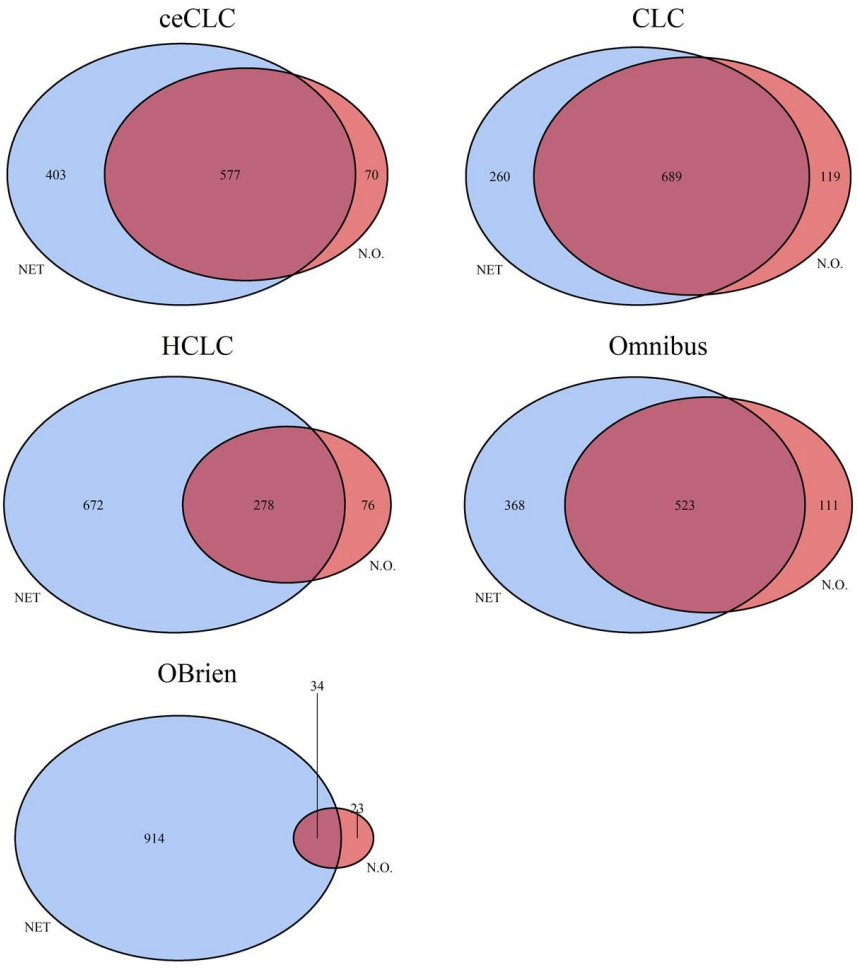


Figure A.17. Tissue expression analysis for mapped genes identified by ceCLC in N.O. (a) and NET (b), respectively.

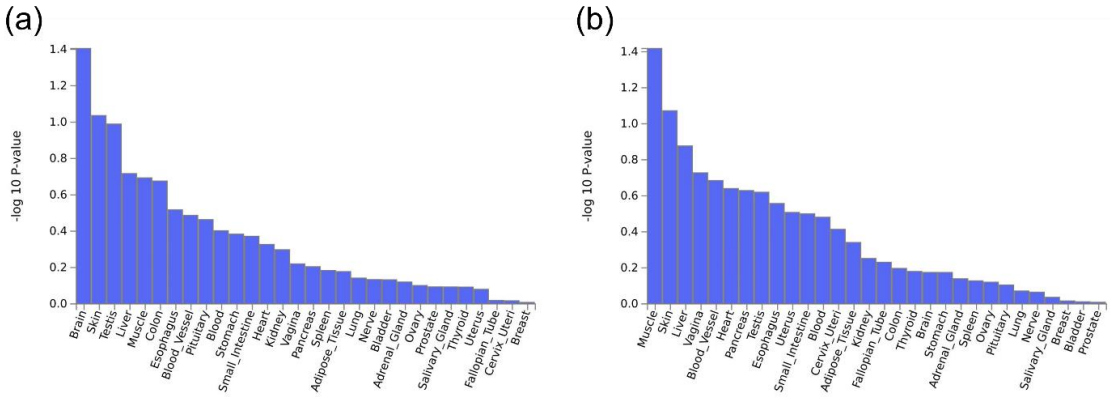
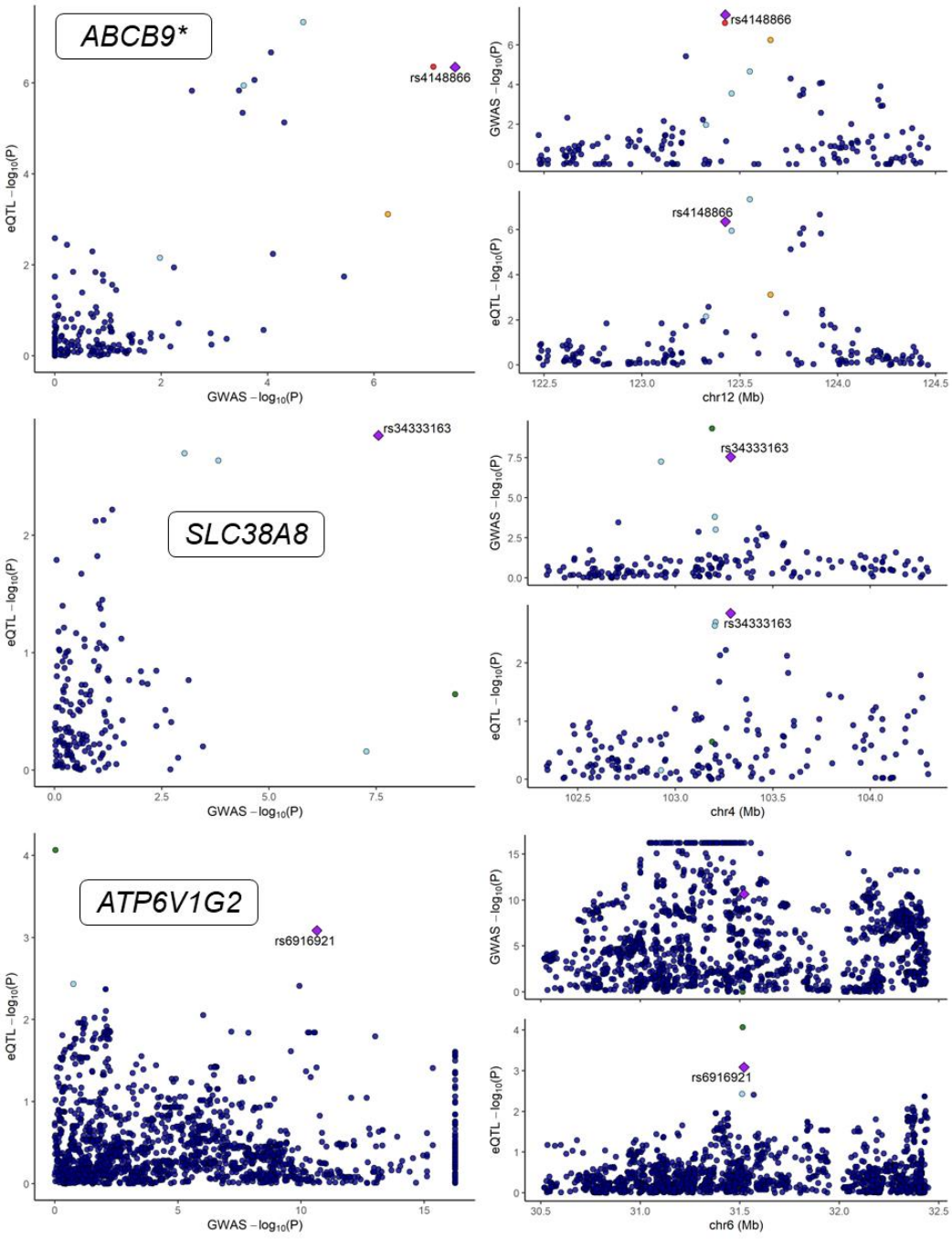


Figure A.18. Colocalization signals. Lead SNPs are selected for colocalization analysis when the top associated SNP identified by ceCLC was also associated with gene expression in the Muscle Skeletal tissue.



B Supplementary Materials for Chapter 2

B.1 Supplementary Texts

Text B.1. Details of other network properties by comparing with random network.

In this study, we also consider two network properties, degree entropy and cross entropy, in comparison between the constructed sparse representation of GPN, $\mathscr{G}_{GPN}^{\tau}$, and the corresponding random network, $\mathscr{G}_{GPN}^{random}$, for a specific $\tau \in (0,1)$.

Degree entropy. The Shannon entropy of the degree can be used to measure the diversity of associations between genetic variants and phenotypes through their degrees[25]. For a specific threshold τ, we define the Shannon entropy for degree of genetic variants and phenotypes as $H_{\tau}^{G} = -\sum_{m=1}^{M} \bar{d}_m^G \log\left(\bar{d}_m^G\right)$ and $H_{\tau}^{P} = -\sum_{k=1}^{K} \bar{d}_k^P \log\left(\bar{d}_k^P\right)$, where the min-max standardized degree are given by $\bar{d}_m^G = \left(d_m^G - \min_m\left\{d_m^G\right\}\right) \Big/ \left(\max_m\left\{d_m^G\right\} - \min_m\left\{d_m^G\right\}\right)$ for the m^{th} genetic variant and $\bar{d}_k^P = \left(d_k^P - \min_k\left\{d_k^P\right\}\right) \Big/ \left(\max_k\left\{d_k^P\right\} - \min_m\left\{d_k^P\right\}\right)$ for the k^{th} phenotype. The global degree entropy of a bipartite network is given by $H_{\tau} = H_{\tau}^{G} + H_{\tau}^{P}$. For the corresponding random network, we use the same way to calculate the degree entropies, $H_{\tau}^{G,random}, H_{\tau}^{P,random}$, and H_{τ}^{random}.

Cross Entropy. We define the cross-entropy of weighted or unweighted degree of genetic variants and phenotypes between $\mathscr{G}_{GPN}^{\tau}$ and $\mathscr{G}_{GPN}^{random}$ to determine the diversity between a bipartite GPN and a random bipartite network.

$$H_{cross}^{G} = -\sum_{m=1}^{M}\left[\bar{d}_m^G \log\left(\bar{d}_m^{G,random}\right) + \left(1-\bar{d}_m^G\right)\log\left(1-\bar{d}_m^{G,random}\right)\right],$$

$$H_{cross}^{P} = -\sum_{k=1}^{K}\left[\bar{d}_k^P \log\left(\bar{d}_k^{P,random}\right) + \left(1-\bar{d}_k^P\right)\log\left(1-\bar{d}_k^{P,random}\right)\right],$$

where, H_{cross}^{G} and H_{cross}^{P} are used to measure the difference between degree distributions of genetic variants and phenotypes in $\mathscr{G}_{GPN}^{\tau}$ and $\mathscr{G}_{GPN}^{random}$, respectively. H_{cross}^{G} and H_{cross}^{P} are always positively valued and it will increase if the degree distributions tend to be more different. Same as degree entropy, we also define the global cross entropy of a bipartite network as $H_{cross} = H_{cross}^{G} + H_{cross}^{P}$. With the loss of the generality, the optimal threshold τ should be selected by maximizing H_{cross}^{G} and H_{cross}^{P}. Meanwhile, in the case of equivalent numbers and weights of edges, the greater the difference of network topologies between $\mathscr{G}_{GPN}^{\tau}$ and $\mathscr{G}_{GPN}^{random}$, the more information the $\mathscr{G}_{GPN}^{\tau}$ includes. Therefore, we also assess the difference of degree entropy between $\mathscr{G}_{GPN}^{\tau}$ and $\mathscr{G}_{GPN}^{random}$ for $\tau \in [0,1]$, which are defined as $\Delta H^{G} = H_{\tau}^{G} - H_{\tau}^{G,random}$ and $\Delta H^{P} = H_{\tau}^{P} - H_{\tau}^{P,random}$. To investigate the significance of differences and the stability of the cross entropies, we construct 1,000 random networks

corresponding to $\mathscr{E}_{GPN}^{\tau}$. For network degree entropy, we evaluate their distributions of the random network, then compare $H_{\tau}^{G}\left(H_{\tau}^{P}\right)$ and the upper bound of the 95% confidence interval (CI) of $H_{\tau}^{G,random}\left(H_{\tau}^{P,random}\right)$. For cross-entropy, we can estimate the standard errors of them and then obtain the stability by computing their 95% CIs.

Text B.2. Details of the five multiple phenotype association tests.

In this study, we apply four powerful GWAS summary-based multiple phenotype association tests to identify the association between phenotypes in each network module and a genetic variant, including minP[26], ACAT[4], MTAG[27], SHom[28]. To simplify the notation, we assume that the tests are applied to test the association between K phenotypes and a genetic variant.

minP. Consider the z-score vector is $\boldsymbol{Z}=\left(Z_1,\cdots,Z_K\right)^T$, where Z_k is the z-score for testing the association between the k^{th} phenotype and a genetic variant. Assume that $\boldsymbol{Z}$ is asymptotically multivariate normal $MVN\left(\boldsymbol{0},\mathbf{R}\right)$ under the null hypothesis that K phenotypes and a genetic variant have no association, where $\mathbf{R}$ is the correlation matrix of phenotypes. The minP test statistic is given by $T_{minP}=\max_{k=1,\cdots,K}\left\{\left|Z_k\right|\right\}$ and the corresponding p-value can be calculated by $p_{minP}=1-\int_{-T_{minP}}^{T_{minP}}\cdots\int_{-T_{minP}}^{T_{minP}} f\left(x_1,\cdots,x_K;\boldsymbol{0},\hat{\mathbf{R}}\right)dx_1\cdots dx_K$, where $f\left(x_1,\cdots,x_K;\boldsymbol{0},\hat{\mathbf{R}}\right)$ is the density function for $MVN\left(\boldsymbol{0},\hat{\mathbf{R}}\right)$ and $\hat{\mathbf{R}}$ can be estimated by using 'estcov' function in *aSPU* package.

ACAT. Let $p_1,\cdots,p_K$ be the p-values of $Z_1,\cdots,Z_K$, respectively. The ACAT test statistic is $T_{ACAT}=\sum_{k=1}^{K}\tan\left\{\left(.5-p_k\right)\pi\right\}/K$ and the p-value of T_{ACAT} is approximated by $p_{ACAT}\approx 0.5-\arctan\left\{T_{ACAT}\right\}/\pi$.

MTAG. Let $\hat{\boldsymbol{\beta}}_{MTAG}=\left(\hat{\beta}_{MTAG,1},\cdots,\hat{\beta}_{MTAG,K}\right)^T$ be the vector of MTAG estimator correcting for both genetic correlation $\boldsymbol{\Omega}$ and phenotypic correlation $\boldsymbol{\Sigma}$ among K phenotypes. $\hat{\boldsymbol{\Omega}}$ can be estimated by the method of moments using the moment condition, and $\hat{\boldsymbol{\Sigma}}$ can be estimated by LD score regression (LDSC)[29]. Let $p_{MTAG,1},\cdots,p_{MTAG,K}$ be the corresponding p-value of MTAG estimator. Then, a Bonferroni correction is used to adjust for multiple testing for K phenotypes.

SHom. The SHom test statistic is given by $T_{SHom}=\left(\boldsymbol{1}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Z}\right)\left(\boldsymbol{1}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Z}\right)^T\Big/\left(\boldsymbol{1}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{1}\right)$, where $\boldsymbol{1}=\left(1,\cdots,1\right)^T$ is a $K\times 1$ vector that contains all 1s and $\hat{\boldsymbol{\Sigma}}$ can be estimated by LDSC. The p-value of T_{SHom} is calculated by assuming T_{SHom} follows a chi-square distribution with 1 degree of freedom.

Text B.3. Simulation studies for PheWAS.

In this study, we expand the application of constructing the bipartite GPN and unipartite PPN to phenome-wide association studies (PheWAS). In PheWAS, correcting for multiple testing is crucial to reduce the risk of false positives and ensure the reliability of the results. Therefore, applying the community detection method for GPN and PPN can obtain a prior grouping of phenotypes based on the shared genetic architecture. Then, jointly testing multiple phenotypes in each network module and one genetic variant can discover the cross-phenotype associations and pleiotropy. Finally, significance thresholds for PheWAS are adjusted for multiple testing by applying the refined false discovery rate (FDR) control approach. We conduct comprehensive simulations to evaluate the FDR of PheWAS based on network modules detected by GPN.

We directly generate a z-score matrix, $\mathbf{Z}$, for M genetic variants and K phenotypes in the whole phenome ($K = 500$ and $1,000$ in our simulation studies). Suppose there are L phenotypic categories and $k = K/L$ in each category. Let $\boldsymbol{\Sigma}$ be the phenotypic correlation matrix, where $\boldsymbol{\Sigma} = Bdiag\left(\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_L\right)$ is a block diagonal matrix, that is, the phenotypes within a category are correlated and between categories are uncorrelated. We consider two scenarios of $\boldsymbol{\Sigma}$[30], SAME and DIFF. In the SAME scenario, there is the same correlation coefficient of each pair of phenotypes within the category, that is, the off-diagonal elements of $\boldsymbol{\Sigma}_l$ equal ρ. In the DIFF scenario, the correlation coefficients are different and $\boldsymbol{\Sigma}_l$ is generated by using an autoregressive (AR(1)) model, that is, $\boldsymbol{\Sigma}_l = \rho^{|k-l|}$. We use $\rho = 0.3$ in the simulation studies.

Assume S_{causal} is the set of M_{causal} true causal variants. Then, we generate a z-score vector for the m^{th} ($m = 1, \cdots, M$) genetic variant from

$$\boldsymbol{Z}_m = \left(Z_{m1}, \cdots, Z_{mK}\right)^T \sim \begin{cases} MVN\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right), & for\ m \notin S_{causal} \\ MVN\left(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}\right), & for\ m \in S_{causal} \end{cases},$$

where $\boldsymbol{\mu}_m = \left(\mu_{m1}, \cdots, \mu_{mK}\right)^T$ is the K dimensional vector of the true effects for causal variants. In the simulation studies, we consider a total of $M = 10^6$ genetic variants and define the first $M_{causal} = 100$ as the true causal variants. Based on the different numbers of phenotypic categories $L = 50$ and 100, we consider the following two models to define $\boldsymbol{\mu}_m$. Let $\boldsymbol{\delta}_1 = \delta\left(1, \cdots, 1\right)^T$ and $\boldsymbol{\delta}_2 = \frac{2\delta}{k/2}\left(1, \cdots, \frac{k}{2}, \frac{k}{2}, \cdots, 1\right)^T$ be two k dimensional vectors of true effect sizes.

Model 1. Only the phenotypes in the first two categories are associated with at least one causal variant with the same effect sizes but in different directions. The is, first 50 causal variants impact the phenotypes in the first category with $\boldsymbol{\delta}_1$; the second 50 causal variants impact the phenotypes in the second category with $-\boldsymbol{\delta}_1$.

Model 2. Only the phenotypes in the first four categories are associated with at least one causal variant with different effect sizes and different directions. That is, the first 25 causal variants impact the phenotypes in the first category with $\boldsymbol{\delta}_1$; the second 25 causal variants impact the phenotypes in the second category with $\boldsymbol{-\delta}_1$; the third 25 causal variants impact the phenotypes in the third category with $\boldsymbol{\delta}_2$; the fourth 25 causal variants impact the phenotypes in the fourth category with $\boldsymbol{-\delta}_2$.

For each simulation model, we run B Monte-Carlo (MC) runs and use the following steps for the b^{th} MC run: i) generate a z-score matrix using the above simulation models; ii) construct the bipartite GPN using the method introduced in section 2.2.1; iii) detect $L^{(b)}$ network modules of K phenotypes using the method introduced in section 2.2.4; iv) test the association between phenotypes in each of $L^{(b)}$ network modules and each and each of M genetic variants using one of the multiple phenotype association tests (**Text B.2**), obtaining $p_{ml}^{(b)}$; v) calculate the optimal threshold, $\hat{p}_m^{(b)}$, by applying the refined FDR controlling approach.

Let $D_m^{(b)} = \sum_{l=1}^{L^{(b)}} \mathrm{I}\left(p_{ml}^{(b)} \le \hat{p}_m^{(b)} \right)$ be the total number of discoveries. Then, we define the true discoveries and false discoveries as $TD_m^{(b)} = \sum_{l \in L_a} \mathrm{I}\left(p_{ml}^{(b)} \le \hat{p}_m^{(b)} \right)$ and $FD_m^{(b)} = D_m^{(b)} - TD_m^{(b)}$, respectively. L_a is the set of network modules containing at least one phenotype that is associated with the m^{th} genetic variant. Therefore, the average FDR can be computed by

$$FDR = \frac{1}{B \times M_{causal}} \sum_{b=1}^{B} \sum_{m=1}^{M_{causal}} \frac{FD_m^{(b)}}{\max\left\{ D_m^{(b)}, 1 \right\}}.$$

Note that we do not generate linkage disequilibrium (LD) of genetic variants in our simulation studies, therefore, we can not use LDSC[29] to estimate the phenotypic correlation matrix $\boldsymbol{\Sigma}$ in applications of SHom. We use the same estimation method introduced in minP and Chisq[26] to approximately estimate $\boldsymbol{\Sigma}$ in the simulation studies. Meanwhile, we directly generate Z-scores instead of effect sizes of genetic variants on phenotypes, therefore, we do not consider MTAG in our simulation studies.

B.2 Supplementary Tables

Table B.1. Phenotypes, abbreviations, samples sizes, disease heritability, and GWAS resources used in heritability enrichment analyses.

Phenotype	Abbreviation	Sample size	Heritability	Reference
Attention deficit/hyperactivity disorder	ADHD	53,293	0.2354 (0.0153)	Demontis et al.[31]
Smoking initiation	SmkInit	632,802	0.0724 (0.0068)	Liu et al.[32]
Autism spectrum disorder	ASD	46,351	0.1941 (0.0168)	Grove et al.[33]
Neuroticism	NSM	170,911	0.0877 (0.0067)	Okbay et al.[34]
Anxiety disorder	AXD	31,890	0.0417 (0.0156)	Meier et al.[35]
Major depressive disorder	MDD	500,199	0.0599 (0.0023)	Howard et al.[36]
Obsessive-compulsive disorder	OCD	9,725	0.3217 (0.0496)	Arnold et al.[37]
Anorexia nervosa	AN	72,517	0.1773 (0.0116)	Watson et al.[38]
Bipolar disorder	BD	51,710	0.3469 (0.0174)	Stahl et al.[39]
Schizophrenia	SCZ	105,318	0.4101 (0.0113)	Pardinas et al.[40]
Educational attainment	EA	766,345	0.1066 (0.0026)	Lee et al.[41]
Cognitive performance	CP	257,828	0.192 (0.0062)	Lee et al.[41]

Notes: Heritability is calculated by LD score regression[29]: heritability (standard error of heritability).

Table B.2. Global genetic correlations (right upper triangle) and proportions of correlated regions (left lower triangle) estimated by SUPERGNOVA[42].

	ADHD	SmkInit	ASD	NSM	AXD	MDD	OCD	AN	BD	SCZ	EA	CP
ADHD		0.47*	0.28*	0.21*	0.45*	0.35*	-0.16	0.008	0.12*	0.07	-0.41*	-0.29*
SmkInit	81%*		0.04	0.15*	0.42*	0.32*	-0.19*	0.01	0.10*	0.13*	-0.35*	-0.15*
ASD	32%*	1.33%		0.25*	0.30*	0.30*	0.10	0.13	0.12	0.19*	0.18*	0.15*
NSM	41%	46%*	15%*		0.41*	0.69*	0.26*	0.26*	0.10	0.18*	-0.24*	-0.17*
AXD	52%	63%*	41%	69%		0.65*	0.15	0.30*	0.20*	0.27*	-0.28*	-0.19*
MDD	69%	65%*	51%*	89%*	77%*		0.20*	0.26*	0.28*	0.29*	-0.17*	-0.10
OCD	0.25%	15%*	0.32%	2.3%*	0.11%*	6.4%*		0.42*	0.23*	0.29*	0.21*	0.01
AN	0.26%	1.52%*	0.19%	54%*	30%*	53%*	37%		0.12*	0.26*	0.20*	0.07
BD	1.37%	2.72%*	4.53%*	1.11%	35%	55%*	2%	21%		0.57*	0.14*	-0.06
SCZ	12%*	39%*	35%*	35%*	42%*	60%*	33%*	58%*	83%*		0.02	-0.22*
EA	76%*	79%*	33%*	61%	51%*	38%*	31%	47%*	32%*	7.5%*		0.63*
CP	67%*	39%*	15%*	38%	20%*	27%*	0.5%	1.9%*	3.4%*	57%*	93%*	

Notes: * indicates the significance genetic correlations and proportions of correlated regions between two phenotypes.

Table B.3. Phenotypic correlations (right upper triangle) and genetic correlations (left lower triangle) estimated by LDSC[29].

	ADHD	SmkInit	ASD	NSM	AXD	MDD	OCD	AN	BD	SCZ	EA	CP
ADHD		0.0034	0.3626	0.0092	0.0040	0.0860	0.0082	-0.1235	0.0398	0.0289	-0.0165	-0.0040
SmkInit	-0.4628*		0.0033	-0.0125	-0.0039	-0.0167	-0.0080	-0.0081	0.0058	0.0055	0.0332	0.0115
ASD	0.3459*	-0.1819*		-0.0125	-0.0024	0.0598	-0.0003	-0.1226	0.0134	0.0170	-0.0003	-0.0013
NSM	0.2642*	-0.1342*	0.2723*		0.0514	0.1211	0.0049	-0.0026	0.0069	0.0036	-0.0403	-0.0212
AXD	0.3008	-0.1579	0.2607	0.9339*		0.0342	0.0027	-0.0108	0.0078	0.0144	-0.0067	-0.0030
MDD	0.4537*	-0.2187*	0.3526*	0.7313*	0.8790*		0.0062	-0.0627	0.0594	0.0375	-0.0146	-0.0111
OCD	-0.1695	0.1185	0.1185	0.2821*	0.2989	0.2591*		-0.0127	0.0376	0.0331	-0.0080	-0.0070
AN	-0.0082	-0.0668	-0.1057	-0.2671*	-0.1976	-0.2870*	-0.4490*		-0.0536	-0.0321	0.0147	0.0116
BD	0.1205	-0.0873	0.1373	0.1213	0.2000	0.3320*	0.3106*	-0.1592*		0.1898	0.0021	0.0119
SCZ	0.1679*	-0.1335*	0.2379*	0.2164*	0.3044*	0.3301*	0.3318*	-0.2527*	0.6667*		-0.0054	-0.0066
EA	-0.5159*	0.2825*	0.2081*	-0.2512*	-0.3416*	-0.1734*	0.2390*	-0.2380*	0.1820*	0.0106		0.1681
CP	-0.3677*	0.0992*	0.2002*	-0.1677*	-0.2459	-0.0866*	0.0456	-0.0819	-0.0701	-0.2438*	0.6840*	

Notes: * indicates the significance genetic correlations between two phenotypes.

Table B.4. Heritability enrichment analyses of network topology annotation (betweenness centrality) calculated from denser and sparse representations of bipartite GPN for each of 12 phenotypes.

Trait	Denser		Sparse ($\tau = 0.45$)		Sparse ($\tau = 0.1$)	
	Enrichment (Standard error) *p-value*	Effect τ^* ($se(\tau^*)$) *z-score*	Enrichment (Standard error) *p-value*	Effect τ^* ($se(\tau^*)$) *z-score*	Enrichment (Standard error) *p-value*	Effect τ^* ($se(\tau^*)$) *z-score*
ADHD	1.1065 (0.0397) *0.0088*	19.9545 (14.6723) *1.3600*	1197.52 (134.831) ***3.18e-23***	10.0116 (0.8860) *11.3003*	498.39 (76.248) ***1.10e-10***	6.9712 (1.0241) *6.8073*
AN	1.1342 (0.0450) ***0.0019***	13.4534 (10.1444) *1.3262*	670.61 (106.029) ***5.40e-11***	4.1649 (0.6006) *6.9350*	292.19 (57.763) ***4.27e-07***	3.1944 (0.6120) *5.2198*
ASD	1.0499 (0.1275) *0.7047*	-10.6735 (34.4418) *-0.3101*	1284.50 (216.018) ***1.47e-16***	7.4179 (0.8219) *9.0256*	129.45 (28.518) ***5.34e-07***	3.9185 (0.7609) *5.1501*
AXD	1.3694 (0.3546) *0.0827*	12.5210 (10.8825) *1.1506*	85.3863 (83.1066) *0.1246*	0.3191 (0.2555) *1.5317*	187.30 (158.517) *0.0387*	0.4038 (0.1944) *2.0767*
BD	1.1634 (0.0258) ***5.94e-14***	40.1394 (6.5713) *6.1082*	1005.12 (109.477) ***3.47e-20***	10.8984 (1.0510) *10.2816*	500.87 (65.368) ***3.70e-14***	10.0078 (1.2369) *8.1479*
CP	1.0888 (0.0143) ***1.94e-09***	9.6528 (4.1834) *2.3074*	863.223 (55.5744) ***7.68e-27***	8.3341 (0.6674) *12.4864*	744.55 (101.854) ***3.04e-08***	7.1091 (1.2339) *5.7613*
EA	1.0876 (0.0096) ***2.20e-17***	6.4808 (1.5985) *4.0543*	991.484 (58.1901) ***5.39e-29***	5.6591 (0.4291) *13.1869*	738.32 (102.370) ***1.31e-07***	4.1130 (0.7518) *5.4712*
MDD	1.1198 (0.0144) ***7.81e-16***	5.2452 (1.0707) *4.8990*	1345.72 (93.838) ***6.24e-25***	2.6991 (0.2275) *11.8632*	624.41 (73.576) ***2.85e-12***	1.9156 (0.2575) *7.4392*
NSM	1.0392 (0.0692) *0.5804*	-2.0694 (11.3005) *-0.1832*	1030.86 (110.868) ***4.37e-16***	3.0760 (0.3472) *8.8586*	730.19 (88.155) ***1.65e-11***	3.4041 (0.4767) *7.1413*
OCD	1.1530 (0.1135) *0.0942*	40.7436 (37.4199) *1.0888*	236.746 (118.689) *0.0141*	3.6566 (1.4789) *2.4725*	52.57 (46.183) *0.2173*	1.0209 (0.8355) *1.2218*
SCZ	1.1640 (0.0198) ***7.51e-16***	60.5467 (10.4817) *5.7764*	1275.95 (86.122) ***3.33e-27***	18.4038 (1.4660) *12.6051*	624.72 (85.240) ***1.95e-09***	13.0659 (2.0790) *6.2848*
SmkInit	1.0719 (0.0221) ***9.78e-05***	3.7899 (1.6834) *2.2514*	568.13 (88.322) ***5.35e-12***	1.8560 (0.2193) *8.4633*	205.23 (45.469) ***1.61e-07***	1.5744 (0.2906) *5.4185*

Notes: The betweenness are scaled by multiplying the number of phenotypes and the number of genetic variants due to it much smaller than the baseline LD annotations. The estimated effect size and its estimated standard error, τ^* and $se(\tau^*)$, are scaled by dividing 10^{-12}. Z-score of the effect size is reported to test the null hypothesis that either $\tau \leq 0$ (one-sided) or $\tau = 0$ (two-sided). P-value of enrichment is reported to test the null hypothesis that $Enrichment > 1$. The bold-faced p-values indicate the annotation is significantly enriched in the disease heritability after accounting for multiple testing (p-value$< 0.05/12 \approx 0.0041$).

Table B.5. The average FDR in the simulation studies for 500 phenotypes and 50 phenotypic categories.

Ce=SAME, Model 1				Ce=SAME, Model 2			
μ	minP	ACAT	SHom	μ	minP	ACAT	SHom
2.0	**0.0846**	0.0522	0.0415	1.5	**0.0860**	0.0462	0.0493
2.2	**0.1203**	0.0632	0.0407	2.0	**0.1111**	0.0583	0.0528
2.5	**0.1040**	0.0630	0.0527	2.5	**0.1001**	0.0512	0.0451
2.8	**0.0959**	0.0532	0.0493	3.0	**0.1017**	0.0521	0.0518
Ce=DIFF, Model 1				Ce=DIFF, Model 2			
μ	minP	ACAT	SHom	μ	minP	ACAT	SHom
1.3	**0.1030**	0.0527	0.0462	1.3	**0.0947**	0.0427	0.0505
1.5	**0.0993**	0.0483	0.0517	1.5	**0.0844**	0.0423	0.0522
1.7	**0.1109**	0.0622	0.0491	1.7	**0.1080**	**0.0645**	0.0453
1.9	**0.0928**	0.0477	0.0482	1.9	**0.1011**	0.0491	0.0427

Notes: FDR is evaluated using 10 MC runs, equivalent to 1,000 replications at a nominal FDR level of 5%. The 95% confidence interval (CI) is [0.0365, 0.0635] and bold-faced values indicate that the values are beyond the upper bounds of the 95% CI.

Table B.6 The average FDR in the simulation studies for 1,000 phenotypes and 100 phenotypic categories.

Ce=SAME, Model 1				Ce=SAME, Model 2			
μ	minP	ACAT	SHom	μ	minP	ACAT	SHom
2.0	**0.0982**	0.0535	0.0496	1.5	**0.0809**	0.0452	0.0473
2.2	**0.1143**	0.0556	0.0482	2.0	**0.1036**	0.0475	0.0500
2.5	**0.1102**	0.0578	0.0501	2.5	**0.1110**	0.0576	0.0556
2.8	**0.1024**	0.0535	0.0481	3.0	**0.1097**	0.0535	0.0526
Ce=DIFF, Model 1				Ce=DIFF, Model 2			
μ	minP	ACAT	SHom	μ	minP	ACAT	SHom
1.3	**0.1068**	0.0535	0.0556	1.3	**0.0997**	0.0425	0.0396
1.5	**0.0925**	0.0503	0.0491	1.5	**0.1002**	0.0560	0.0431
1.7	**0.0942**	0.0460	0.0483	1.7	**0.0914**	0.0412	0.0483
1.9	**0.0977**	0.0470	0.0483	1.9	**0.0998**	0.0555	0.0515

Notes: FDR is evaluated using 10 MC runs, equivalent to 1,000 replications at a nominal FDR level of 5%. The 95% confidence interval (CI) is [0.0365, 0.0635] and bold-faced values indicate that the values are beyond the upper bounds of the 95% CI.

B.3 Supplementary Figures

Figure B.1. Network connectance of GPN with different thresholds for (a) 12 genetically correlated phenotypes and (b) 588 EHR-derived phenotypes in the UK Biobank.

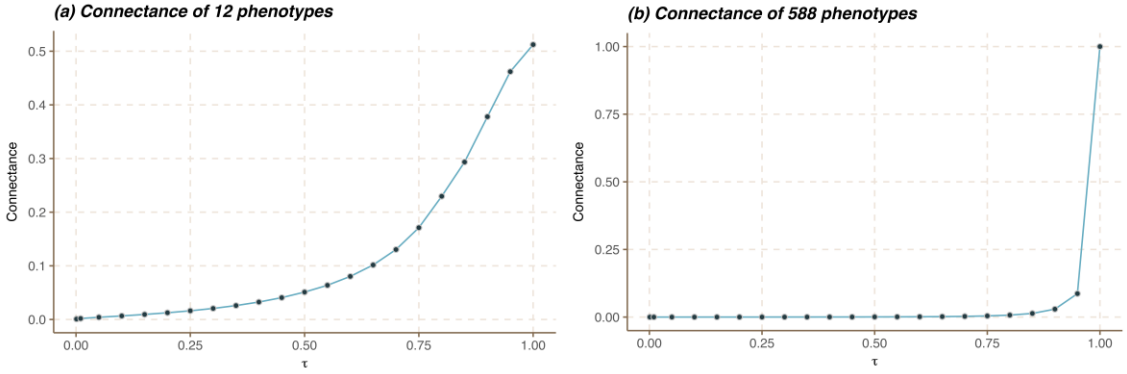


Figure B.2. Network properties of the unweighted bipartite GPNs for 12 genetically correlated phenotypes. (a) KL divergency for genetic variants. The blue line is the mean of KL divergencies across 1,000 random network comparisons. The boxplots show the scaled distribution of KL divergency for each threshold. (b) Cross entropy for genetic variants. Blue lines are the means of the cross entropy across 1,000 random network comparisons. The boxplots show the scaled distribution of cross entropy for each threshold. Red lines represent the degree entropy for the original network. The boxplots show the distribution of degree entropy for each threshold across 1,000 random networks. The blue line represents the difference between the original and random networks. (c) Unweighted degree entropy for genetic variants. (d) plot of the unweighted degree distribution of genetic variants for three GPNs on the log-log scale, denser representation ($\tau = 1$), well-defined sparse representation ($\tau = 0.45$), and an arbitrary threshold sparse representation ($\tau = 0.1$).

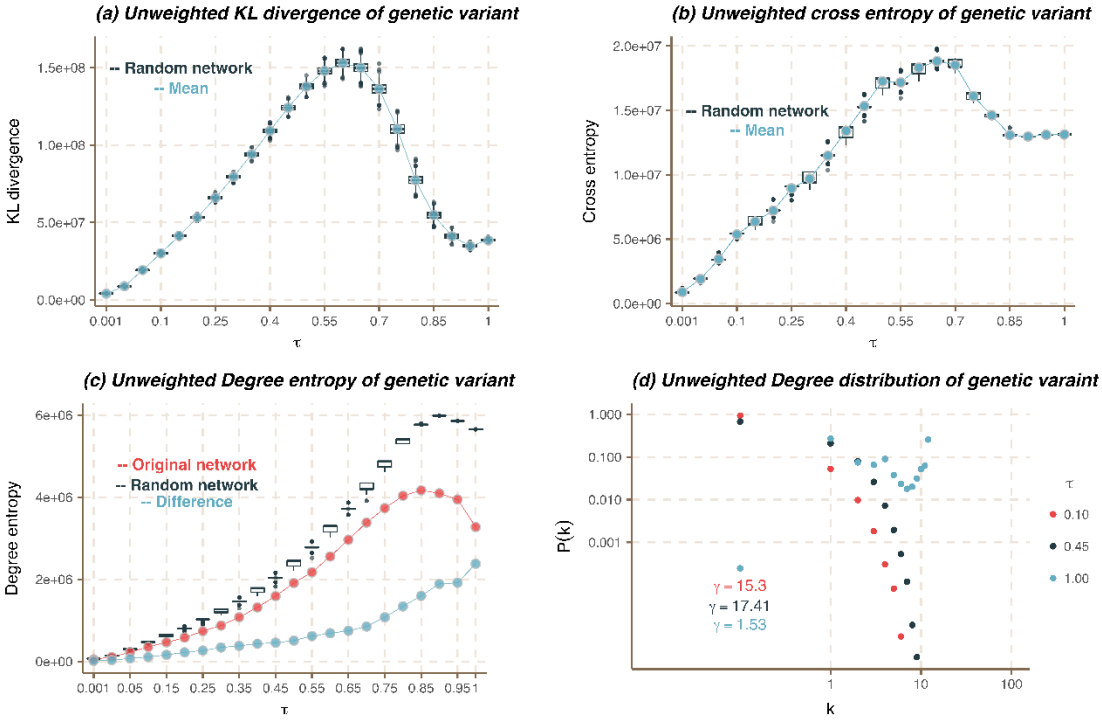


Figure B.3. The correlation of 12 highly correlated phenotypes calculated by different methods: (a) Adjacency matrix of Phenotype and Phenotype Network (PPN) projected from the denser representation of the bipartite GPN; (b) Adjacency matrix of PPN from the well-defined sparse representation of GPN; (c) Genetic correlation matrix estimated by LDSC; (d) Global genetic correlation estimated by SUPERGNOVA.

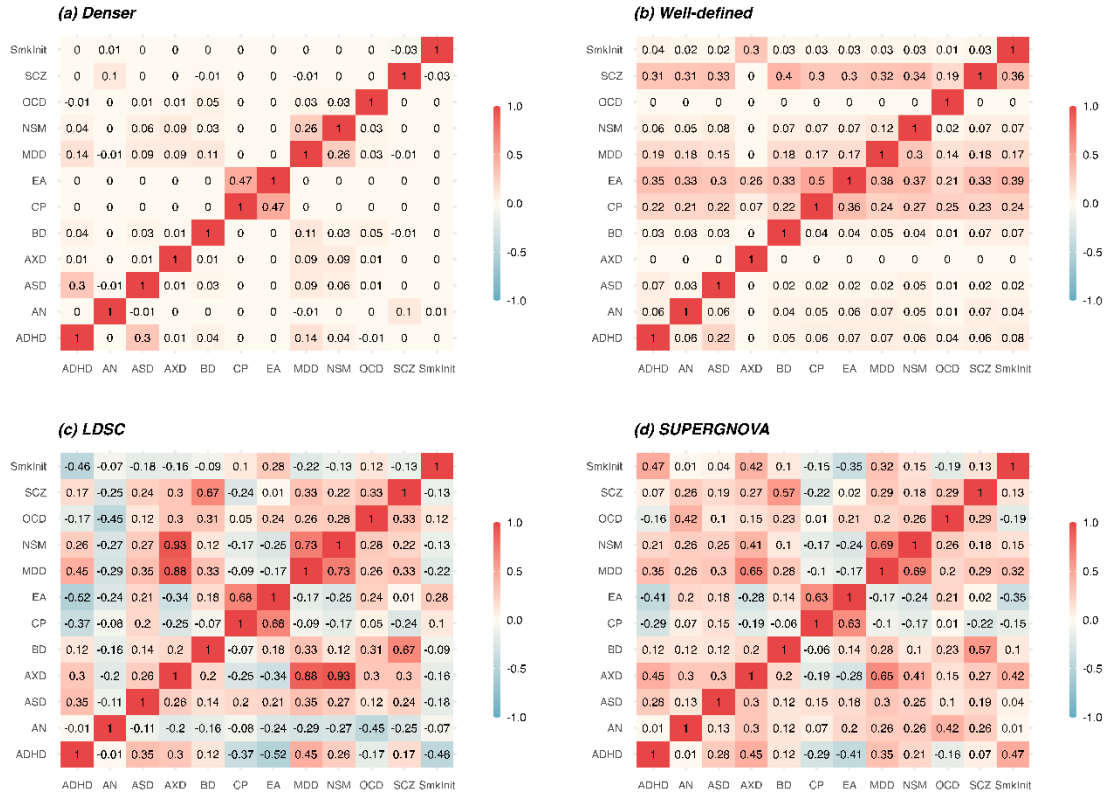


Figure B.4. The qq-plot of EA versus CP based on weight of (a) the denser representation of GPN and (b) the well-defined sparse representation of GPN. The qq-plot of EA versus CP based on (c) -log10(p-values) and (d) z-scores from GWAS summaries.

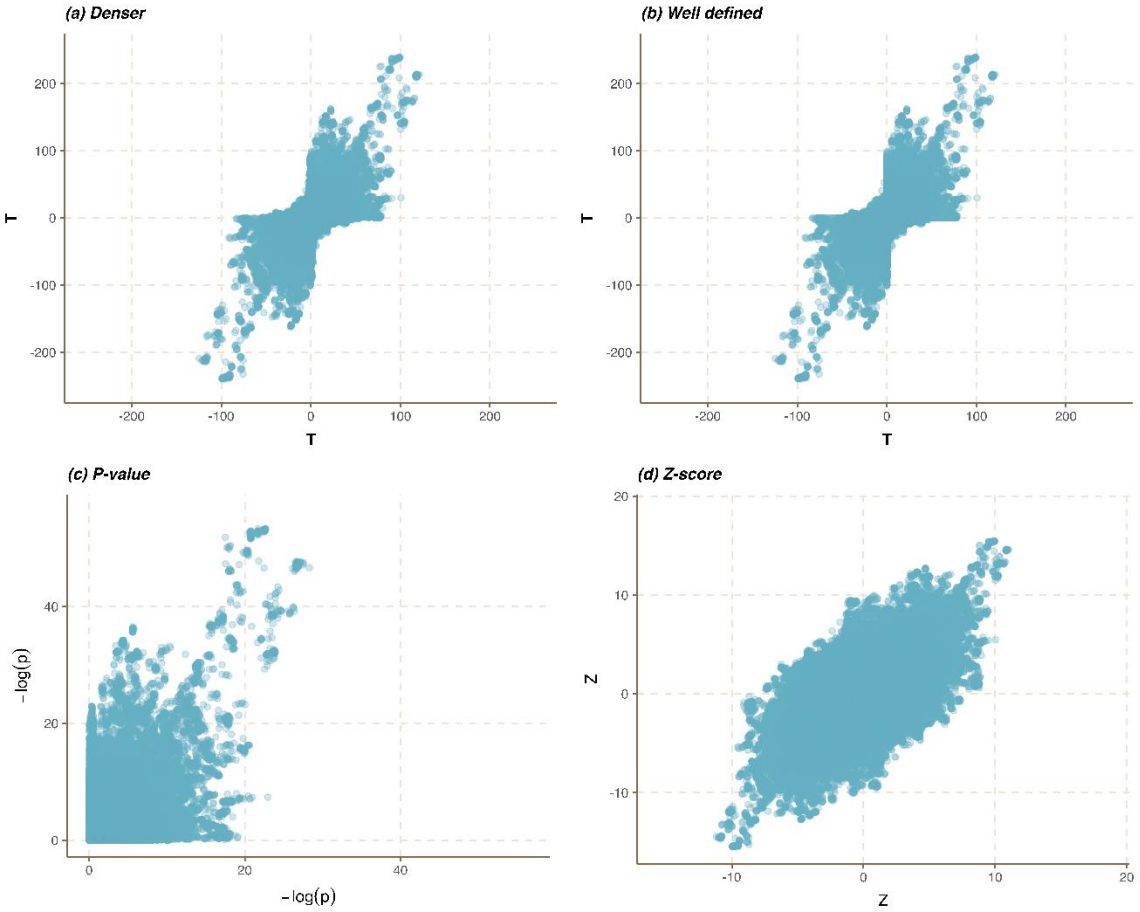


Figure B.5. Heatmap of edge weights in the well-defined sparse representation of GPN for (a) the top 100 and (b) the top 1000 genetic variants with the highest degree centrality.

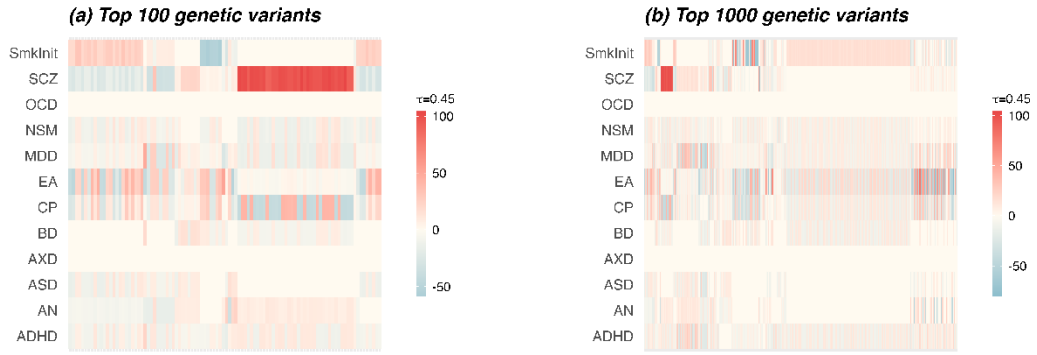


Figure B.6. Cross entropy and degree entropy the unweighted bipartite GPNs for 588 EHR-derived phenotypes in the UK Biobank. Cross entropy for (a) genetic variants and (b) phenotypes. The blue line is the mean of cross entropy across 1,000 random network comparisons. The boxplots show the scaled distribution of cross entropy for each threshold. Degree entropy for (a) genetic variants and (b) phenotypes. Red lines represent the degree entropy for the original network. The boxplots show the distribution of degree entropy for each threshold across 1,000 random networks. The blue line represents the difference between the original and random networks.

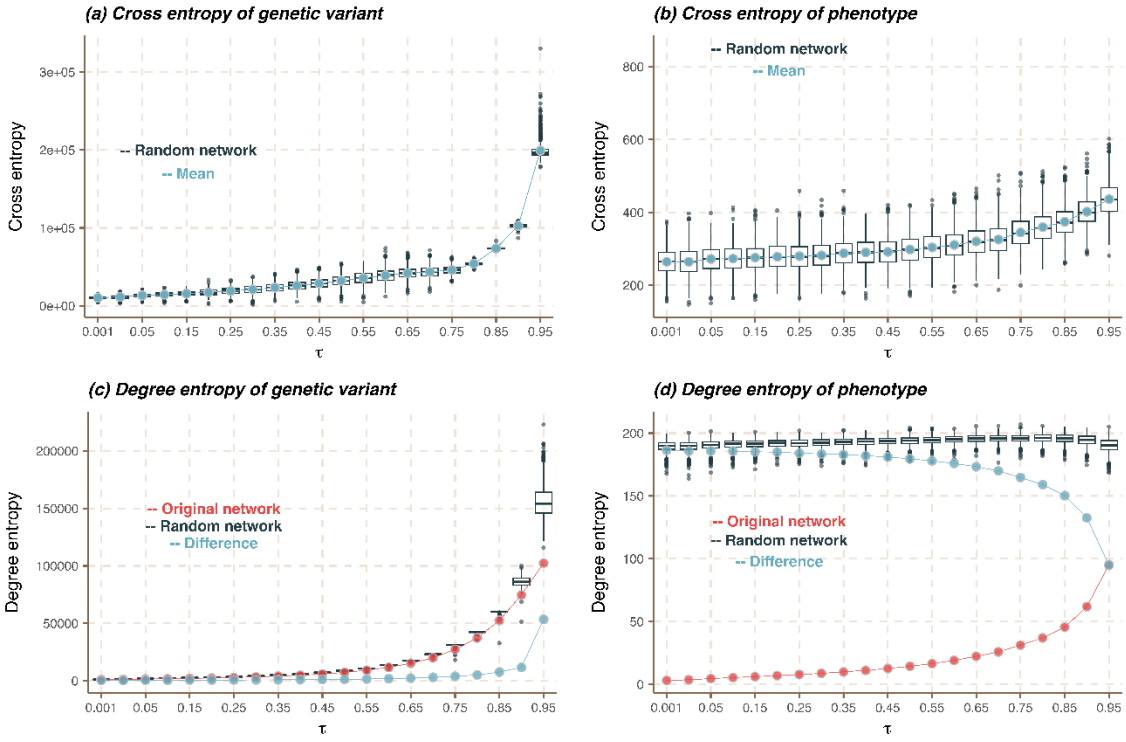


Figure B.7. Network properties of the unweighted bipartite GPNs for 588 EHR-derived phenotypes in the UK Biobank. (a) and (b) KL divergency for genetic variants and phenotypes. The blue line is the mean of KL divergencies across 1,000 random network comparisons. The boxplots show the scaled distribution of KL divergency for each threshold. (c) and (d) Cross entropy for genetic variants and phenotypes. The blue line is the mean of cross entropy across 1,000 random network comparisons. The boxplots show the scaled distribution of cross entropy for each threshold. (e) and (f) Unweighted degree entropy for genetic variants and phenotypes. The red line represents the degree entropy for the original network. The boxplots show the distribution of degree entropy for each threshold across 1,000 random networks. The blue line represents the difference between the original and random networks. (g) and (h) Unweighted degree distribution of genetic variants and phenotypes for four GPNs, more denser representation ($\tau = 0.8$), well-defined sparse representation ($\tau = 0.6$), and two arbitrary threshold sparse representations ($\tau = 0.2$ and $\tau = 0.4$).

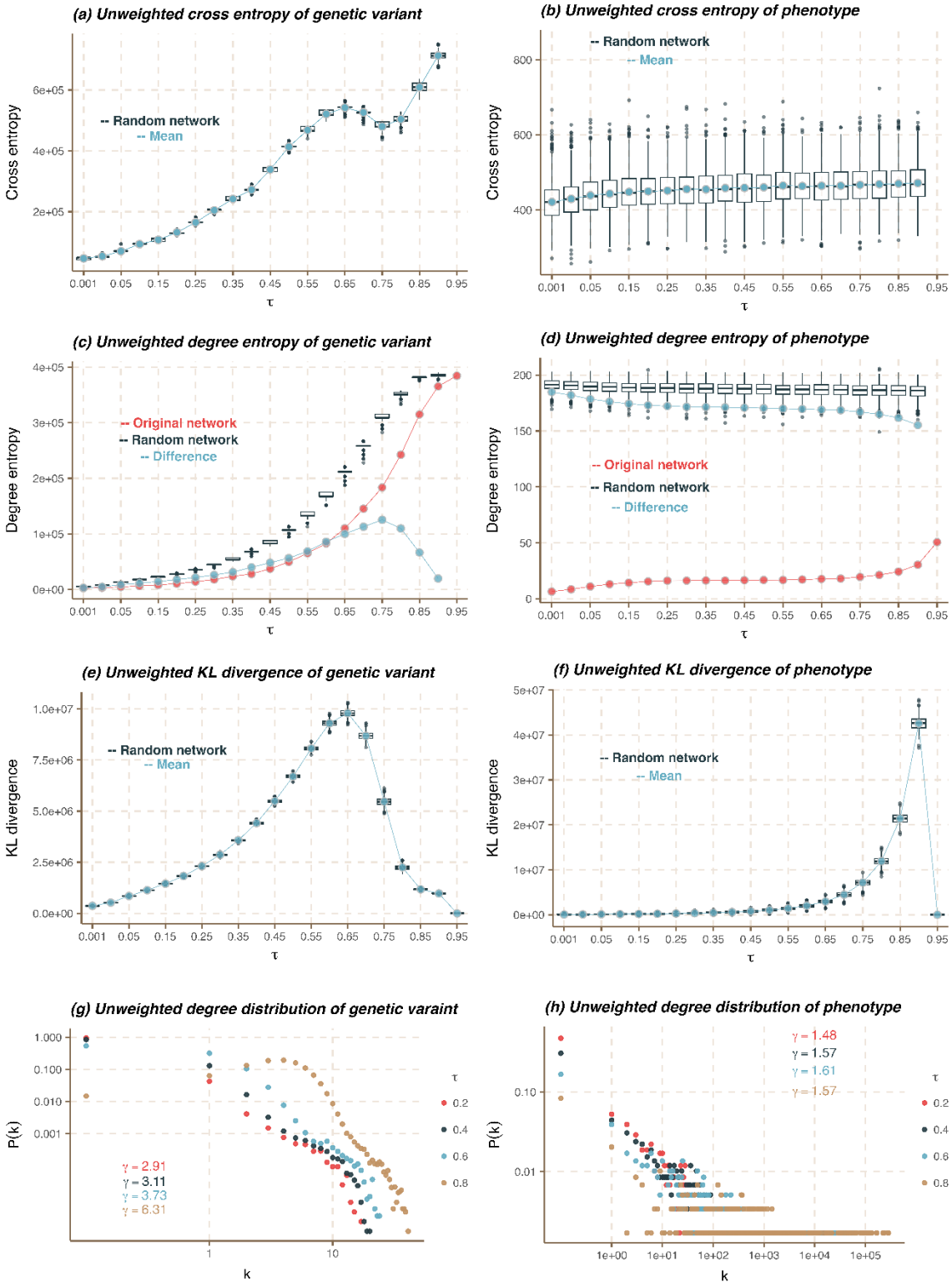

(a) Unweighted cross entropy of genetic variant
Cross entropy
-- Random network
-- Mean
τ
(b) Unweighted cross entropy of phenotype
Cross entropy
-- Random network
-- Mean
τ
(c) Unweighted degree entropy of genetic variant
Degree entropy
-- Original network
-- Random network
-- Difference
τ
(d) Unweighted degree entropy of phenotype
Degree entropy
-- Original network
-- Random network
-- Difference
τ
(e) Unweighted KL divergence of genetic variant
KL divergence
-- Random network
-- Mean
τ
(f) Unweighted KL divergence of phenotype
KL divergence
-- Random network
-- Mean
τ
(g) Unweighted degree distribution of genetic varaint
P(k)
γ = 2.91
γ = 3.11
γ = 3.73
γ = 6.31
τ
0.2
0.4
0.6
0.8
k
(h) Unweighted degree distribution of phenotype
P(k)
γ = 1.48
γ = 1.57
γ = 1.61
γ = 1.57
τ
0.2
0.4
0.6
0.8
k

Figure B.8. Degree centrality and betweenness centrality of genetic variants of the weighted bipartite GPNs for 588 EHR-derived phenotypes in the UK Biobank.

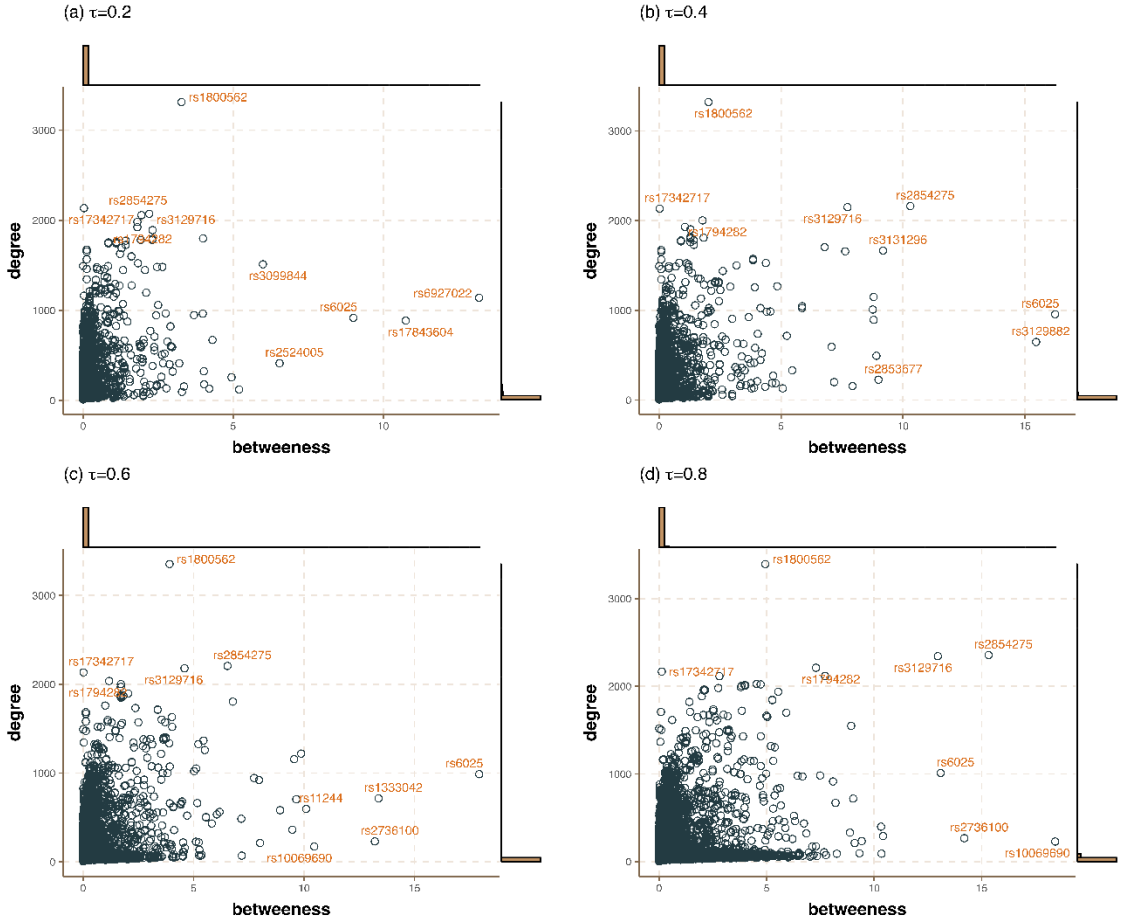


Figure B.9. Degree centrality and betweenness centrality of genetic variants of the unweighted bipartite GPNs for 588 EHR-derived phenotypes in the UK Biobank.

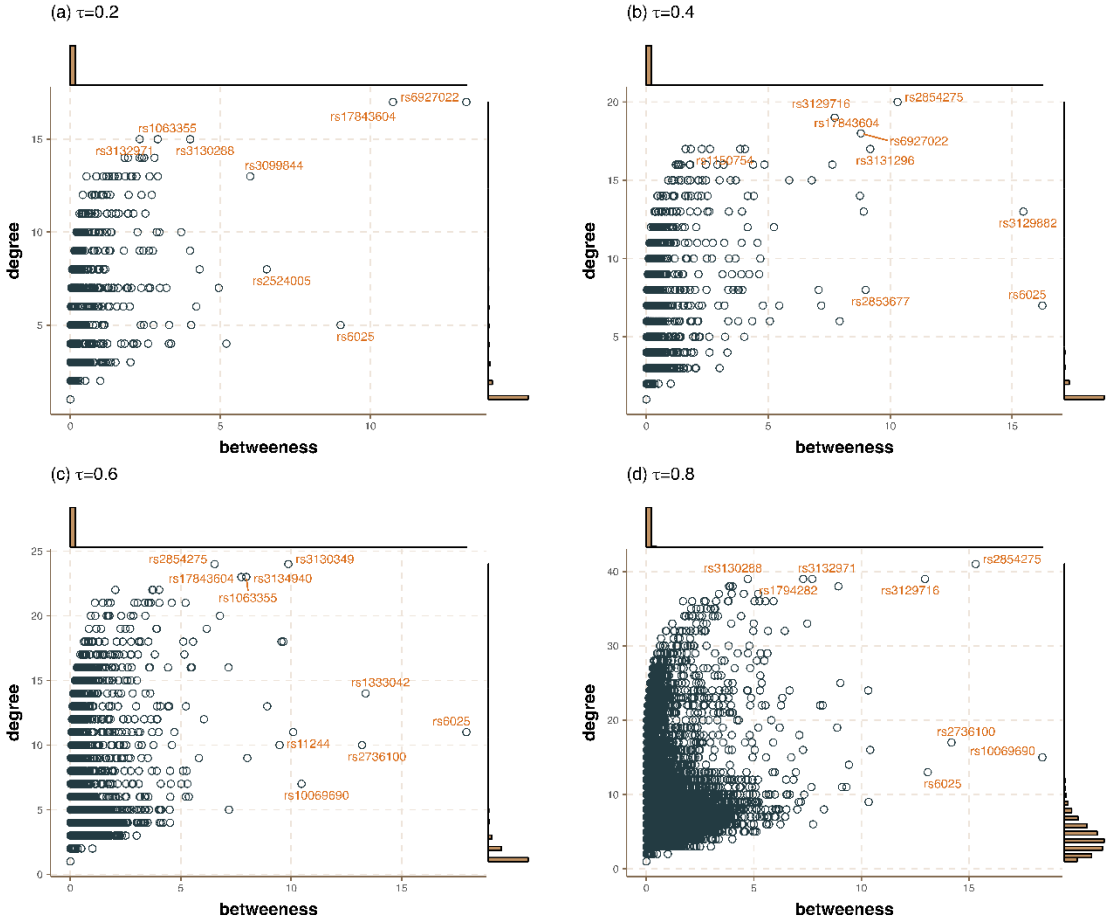
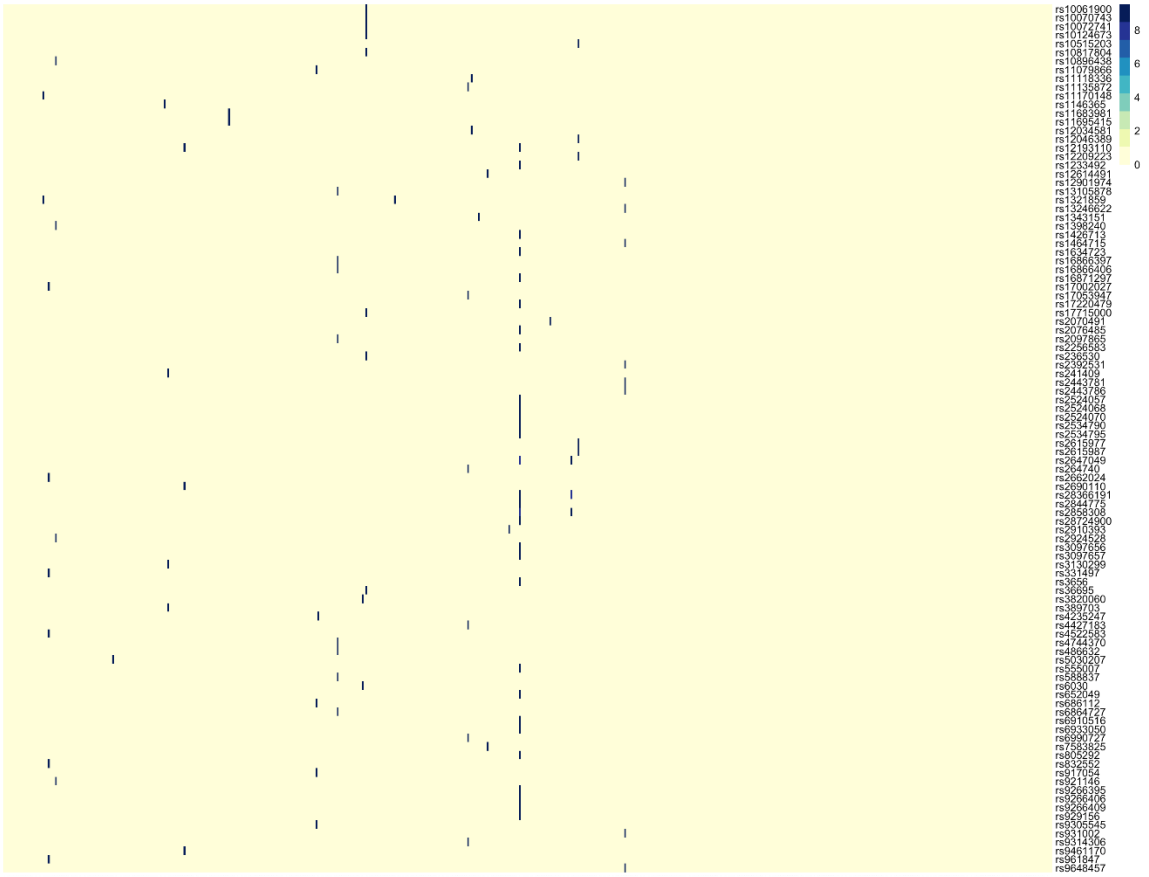


Figure B.10. Heatmap of -$\log_{10}$(p-value) of 100 genetic variants from GWAS summary datasets, which are uniquely identified by ACAT based on LDSC compared with ACAT based on the UK Biobank. Only the p-values smaller than GWAS significance level (5×10^{-8}) shown in the heatmap.



C Supplementary Materials for Chapter 3

C.1 Supplementary Tables

Table C.1. Information of the 18 genes used to obtain the number of replications in the estimation of $\boldsymbol{\Omega}$ and to evaluate type I error rates of Overall.

gene	position	# SNPs	Average LD
AGTRAP	chr1: 11736084 - 11754802	23	12.72
TP53	chr17: 7661779 - 7687538	25	13.78
OR8D2	chr11:124319262-124320197	27	65.44
FNBP4	chr11:47716494-47767443	29	108.34
HLA-DOA	chr6: 33004182 - 33009591	38	13.21
C3orf22	chr3: 126526999 - 126558965	40	24.29
GOSR1	chr17: 17:30477362-30527592	40	78.42
LRRFIP2	chr3:37052626-37183689	56	92.67
MCU	chr10:72692131-72887694	56	144.73
C11orf49	chr11:46936689-47164385	79	126.56
HLA-DOB	chr6: 32812763 - 32820466	85	19.34
AKR1E2	chr10: 4786629 - 4848062	89	16.05
DOCK3	chr3:50674927-51384198	102	170.85
CCDC7	chr10:32446140-32882874	117	56.56
SYNE2	chr14: 63761899 - 64226433	174	36.13
UGT1A10	chr2: 233636454 - 233773305	189	38.69
MCPH1	chr8: 6406596 - 6648508	262	20.86
CDH13	chr16: 82626965 - 83800640	359	17.25

Notes: "# SNPs" indicates the number of SNPs in the corresponding gene. "Average LD" indicates the average of LD scores of SNPs in the gene.

Table C.2. Estimated type I error rates of Overall divided by the significance level for each of the 17 genes at different significance levels with 2×10^5 replications.

Gene	1×10^{-2}	1×10^{-3}	1×10^{-4}
TP53	1.03	1.23	1.12
OR8D2	0.73	0.78	0.80
FNBP4	0.94	1.08	1.15
HLA-DOA	0.97	1.12	1.10
C3orf22	0.98	1.08	0.95
GOSR1	0.95	1.01	0.95
LRRFIP2	0.98	1.04	1.00
MCU	0.81	0.85	0.80
C11orf49	0.96	0.93	1.05
HLA-DOB	0.91	1.07	1.19
AKR1E2	1.10	1.07	1.21
DOCK3	0.98	0.94	0.65
CCDC7	1.01	1.20	1.10
SYNE2	0.94	1.03	1.00
UGT1A10	0.97	1.03	0.86
MCPH1	0.99	1.17	1.14
CDH13	1.10	1.13	1.12

C.2 Supplementary Figures

Figure C.1. The p-values to test if the estimated correlation matrix of p-values based on $B=10^4$ and the estimated correlation matrix of p-values based on B_0 are the same for the 18 genes. The red dotted line indicates the significant level 0.05.

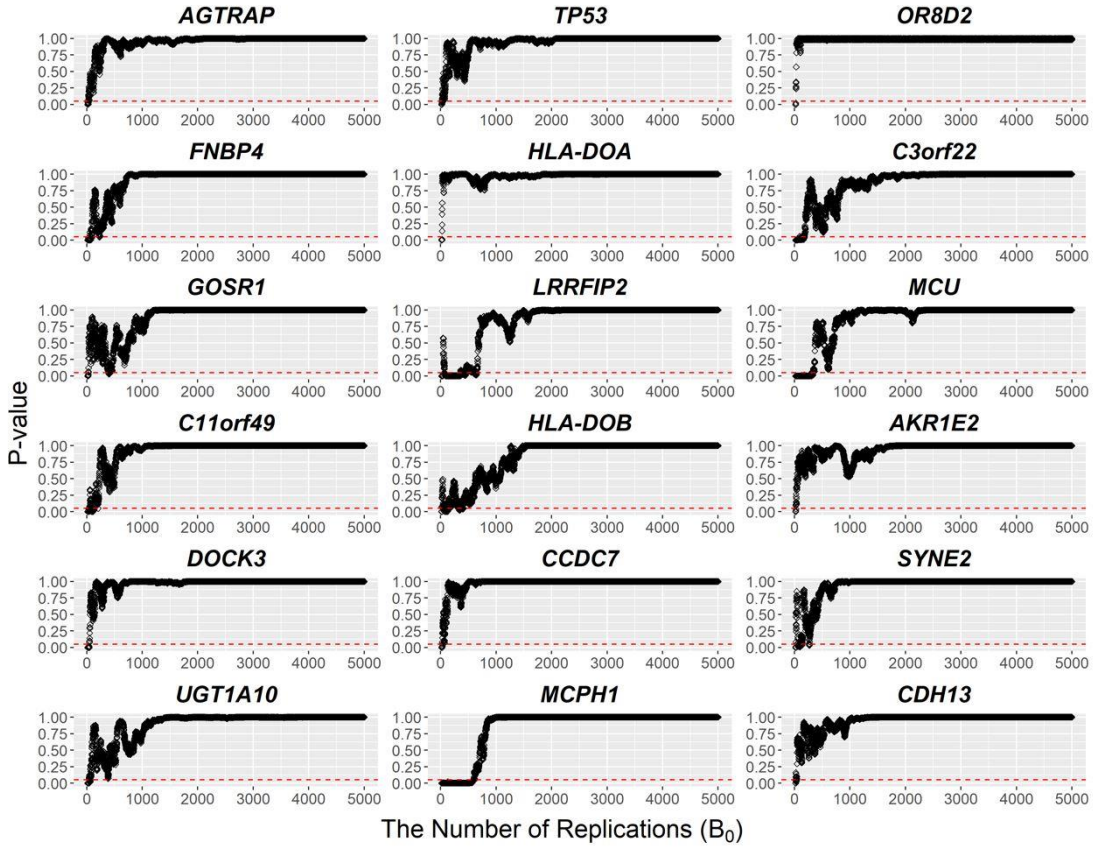


Figure C.2. The LD block structures of gene *AGTRAP* (left) and gene *C3orf22* (right).

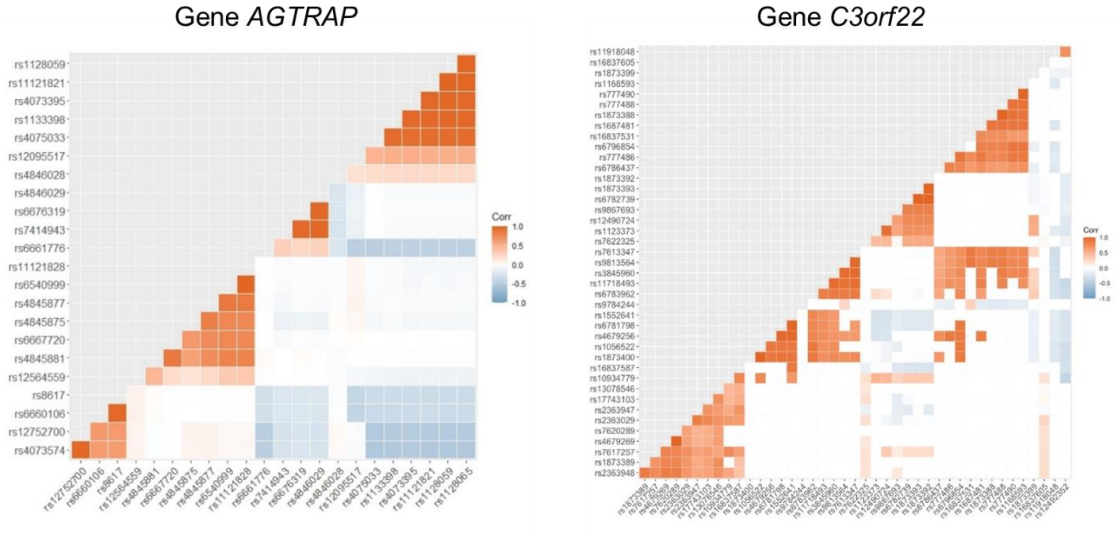


Figure C.3. Estimated correlation matrix of p-values $\hat{\boldsymbol{\Omega}}$ for gene *AGTRAP*.

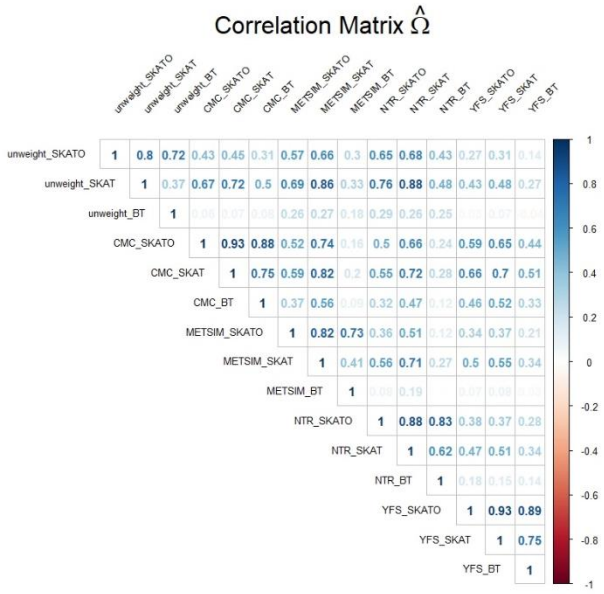


Figure C.4. Power comparisons of gene-based association tests at 1.75×10^{-6} significance level under Uni-directional effects ($\beta_1=\beta_2=\beta_3=\beta_4$) with $p_{causal}=(0.1,0.2,0.3,0.4)$ based on gene *C3orf22*. (a) Estimated power against phenotypic heritability h_p^2 with fixed expression heritability $h_e^2=0.2$; (b) Estimated power against expression heritability h_e^2 with fixed phenotypic heritability $h_p^2=0.2$.

(a) $h_e^2=0.2$ and $h_p^2=0.2,0.4,0.6,0.8$ (b) $h_p^2=0.2$ and $h_e^2=0.2,0.4,0.6,0.8$

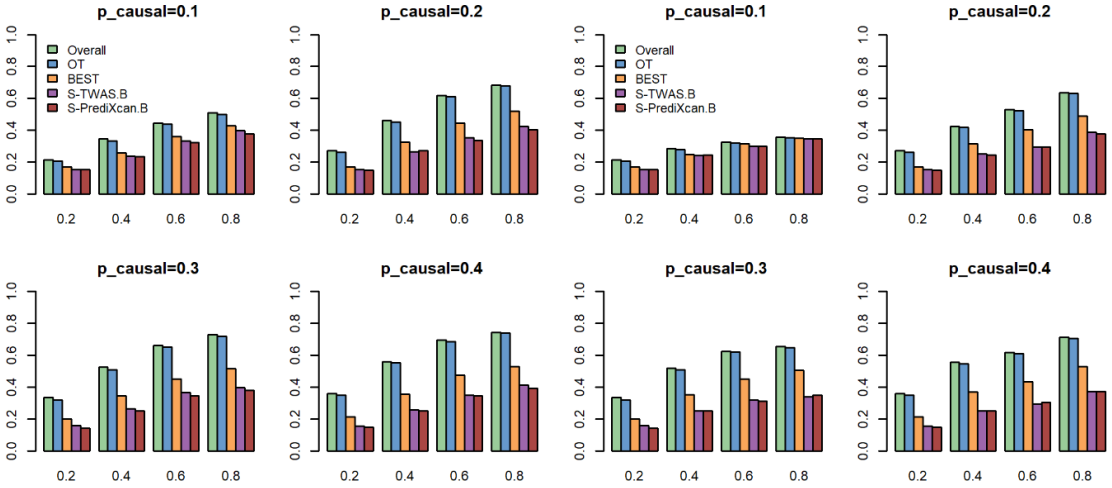


Figure C.5. Power comparisons of gene-based association tests at 1.75×10^{-6} significance level under Bi-directional effects ($\beta_1=\beta_2=-\beta_3=-\beta_4$) with $p_{causal}=(0.1,0.2,0.3,0.4)$ based on gene *C3orf22*. (a) Estimated power against phenotypic heritability h_p^2 with expression heritability $h_e^2=0.2$; (b) Estimated power against expression heritability h_e^2 with phenotypic heritability $h_p^2=0.2$.

(a) $h_e^2=0.2$ and $h_p^2=0.2,0.4,0.6,0.8$ (b) $h_p^2=0.2$ and $h_e^2=0.2,0.4,0.6,0.8$

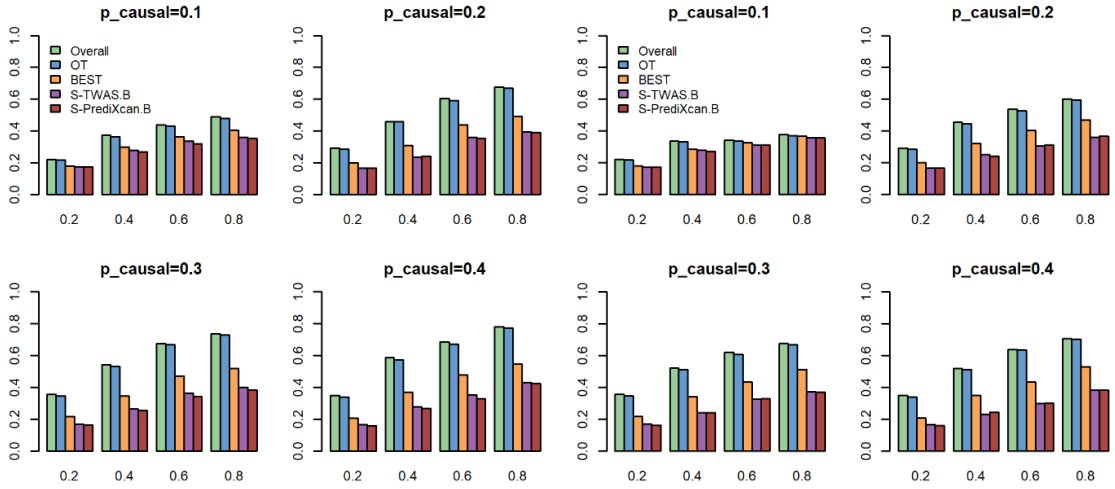


Figure C.6. Power comparisons of gene-based association tests at 1.75×10^{-6} significance level with $p_{causal}=(0.2,0.3)$ based on gene *C3orf22* with eQTL - derived weights from $K=20$ studies. Estimated power against phenotypic heritability h_p^2 with expression heritability $h_e^2=0.2$. (a) Uni-directional effects ($\beta_1=\cdots=\beta_K$); (b) Bi-directional effects ($\beta_1=\cdots=\beta_{K/2}=-\beta_{K/2+1}=\cdots=-\beta_K$).

(a) Uni-directional effects (b) Bi-directional effects

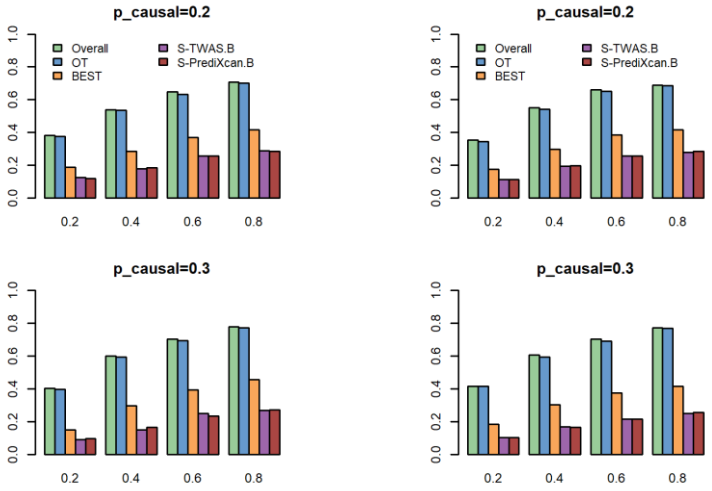


Figure C.7. Power comparisons of gene-based association tests at 1.75×10^{-6} significance level under Uni-directional effects ($\beta_1=\beta_2=\beta_3=\beta_4$) with noise to the eQTL for $p_{causal}=(0.1,0.2,0.3,0.4)$ based on gene *C3orf22*. (a) Estimated power against phenotypic heritability h_p^2 with expression heritability $h_e^2=0.2$; (b) Estimated power against expression heritability h_e^2 with phenotypic heritability $h_p^2=0.2$.

(a) $h_e^2=0.2$ and $h_p^2=0.2,0.4,0.6,0.8$ (b) $h_p^2=0.2$ and $h_e^2=0.2,0.4,0.6,0.8$

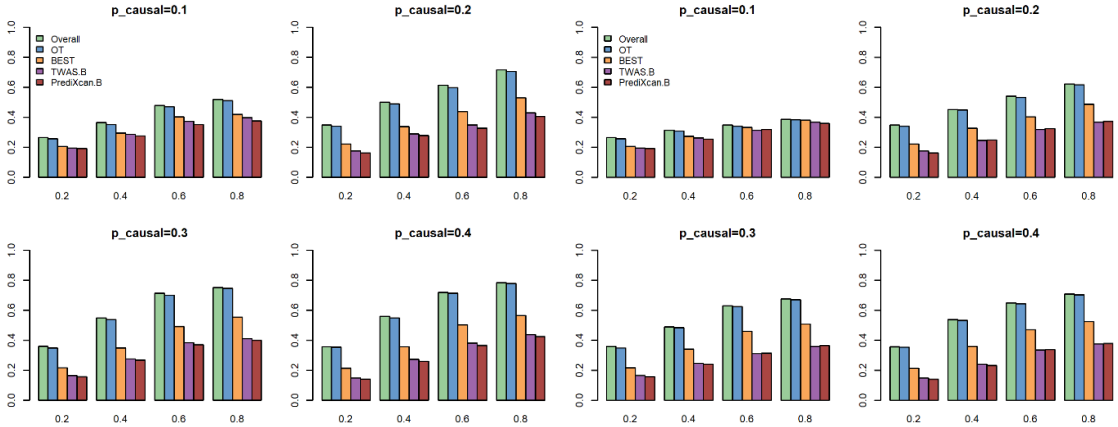


Figure C.8. Power comparisons of gene-based association tests at 1.75×10^{-6} significance level under Bi-directional effects ($\beta_1=\beta_2=-\beta_3=-\beta_4$) with noise to the eQTL for $p_{causal}=(0.1,0.2,0.3,0.4)$ based on gene *C3orf22*. (a) Estimated power against phenotypic heritability h_p^2 with expression heritability $h_e^2=0.2$; (b) Estimated power against expression heritability h_e^2 with phenotypic heritability $h_p^2=0.2$.

(a) $h_e^2=0.2$ and $h_p^2=0.2,0.4,0.6,0.8$ (b) $h_p^2=0.2$ and $h_e^2=0.2,0.4,0.6,0.8$

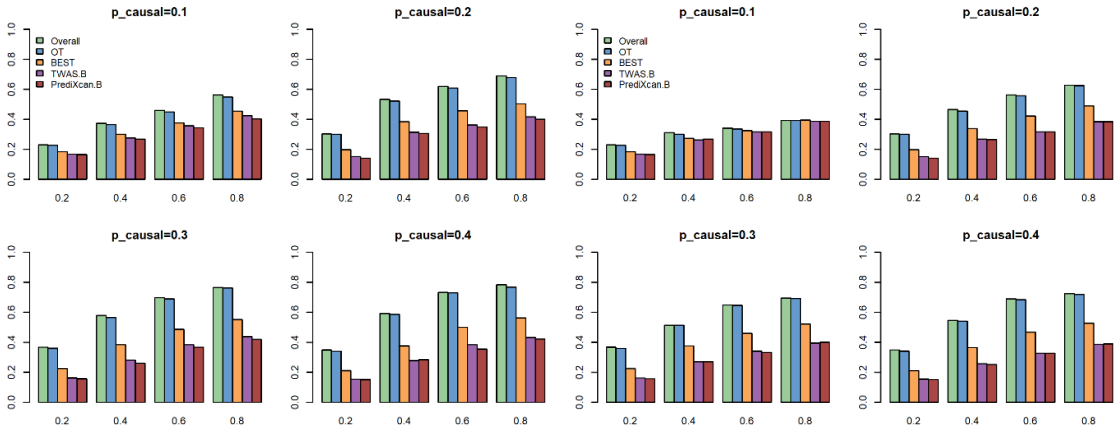


Figure C.9. Estimated power against phenotypic heritability h_p^2 with expression heritability $h_e^2 = 0.05$ at 1.75×10^{-6} significance level on gene *C3orf22* with $p_{causal} = (0.1, 0.2)$ and sample size of 100,000. (a) Uni-directional effects with noise to eQTL; (b) Bi-directional effects with noise to eQTL.

(a) Uni-directional effects (b) Bi-directional effects

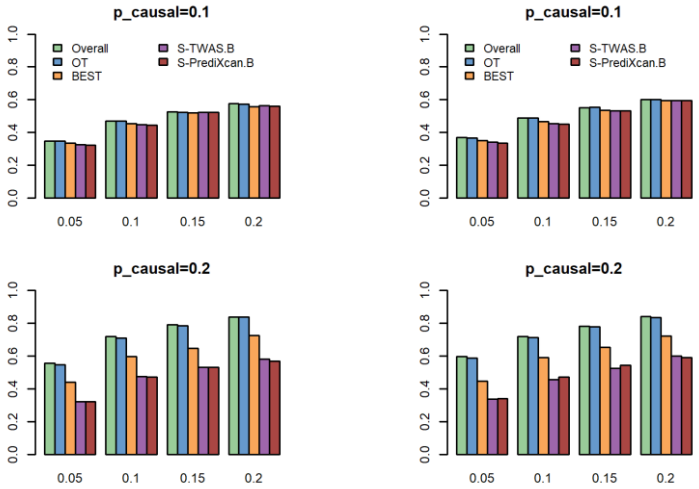
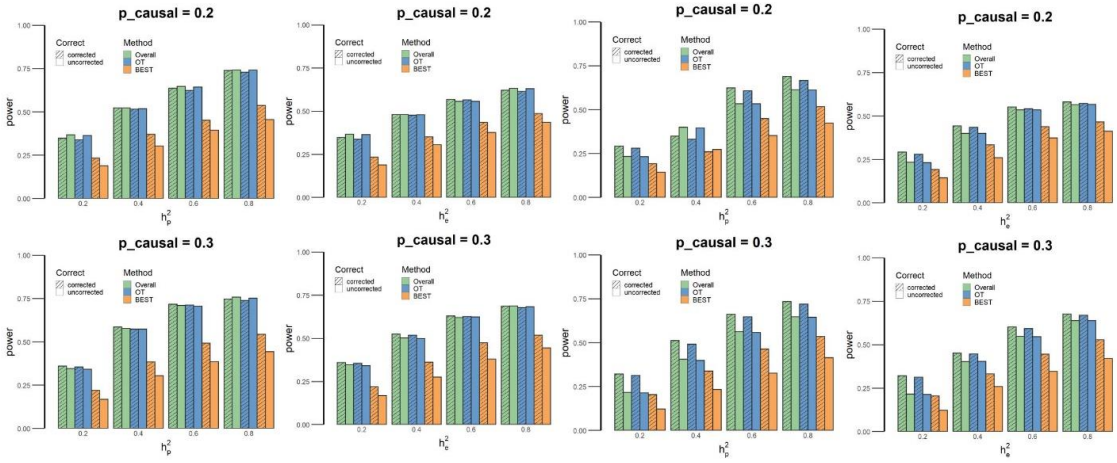


Figure C.10. Power comparisons of Overall, OT, and BEST based on corrected and uncorrected LD structure on gene C3orf22 with $p_{causal} = (0.2, 0.3)$ at 1.75×10^{-6} significance level. (a) and (b): Uni-directional effects with noise to eQTL; (b) and (d): Bi-directional effects with noise to eQTL.

(a) $h_e^2 = 0.2$ (b) $h_p^2 = 0.2$ (c) $h_e^2 = 0.2$ (d) $h_p^2 = 0.2$



D Supplementary Materials for Chapter 4

D.1 Supplementary Texts

Text D.1. The general simulation setting.

We considered the general simulation setting for comparison. To simulate expression levels of p target genes (TGs), we used the following linear model,

$$\boldsymbol{x}_i = y_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i,$$

Here $\boldsymbol{x}_i = \left(x_{i1}, \cdots, x_{ip}\right)^T$ represents the expression level of p TGs in sample i. y_i is the expression level of a transcription factor (TF) in sample i and was generated from a standard normal distribution. $\boldsymbol{\beta} = \left(\beta_1, \cdots, \beta_p\right)^T$ represents the fixed regulation effects of the TF on p TGs. $\boldsymbol{\varepsilon}_i$ represents the error terms for p TGs in sample i, where $\boldsymbol{\varepsilon}_i$ was generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}_p$ (identity matrix), $\boldsymbol{\varepsilon}_i \sim MVN_p(\mathbf{0}, \boldsymbol{I}_p)$. We used $n = 300$ samples, $p = 500$ TGs in this simulation studies.

The regulation effects $\boldsymbol{\beta}$ were determined based on the relationship between TGs and the TF. In the general simulation settings, only the first 50 TGs were regulated by the TF. Therefore, the regulation effects $\boldsymbol{\beta}$ were defined as

$$\beta_j = \begin{cases} \beta, & if\ j \in (1, \cdots, 25), \\ -\beta, & if\ j \in (25, \cdots, 50), \\ 0, & otherwise. \end{cases}$$

Text D.2. Simulation settings if the target genes have the biological network structure.

To simulate correlated expression levels of p TGs within a biological network, we added the network factor into the general linear model,

$$\boldsymbol{x}_i = y_i\boldsymbol{\beta} + \boldsymbol{Z}_i + \boldsymbol{\varepsilon}_i,$$

Here $\boldsymbol{Z}_i \sim MVN_p(\mathbf{0}, \boldsymbol{\Sigma})$ is the network factor values in sample i with a network structure, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{x}_i$, and $\boldsymbol{Z}_i$ was generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. For a given network, $\boldsymbol{\Sigma}$ was simulated by the following ways, as described by Peng et al.[43] and Cao et al.[44]. First, an initial concentration matrix is generated. For a pair of TGs m and k ($m = 1, \cdots, p, k = 1, \cdots, p$), the corresponding element in the initial concentration matrix was set as 0 if they were not linked or was generated from a uniform distribution on $[-0.7, -0.1] \cup [0.1, 0.7]$ if they were linked. Then the non-zero elements in the initial concentration matrix were rescaled to assure its positive definiteness and the rescaled matrix was averaged with its transpose to ensure the symmetry. Denote $W = (\omega_{mk})$ as the inverse of the matrix after rescaling and averaging based on the initial concentration matrix, the element Σ_{jk} in the covariance matrix $\boldsymbol{\Sigma}$ was determined by $\Sigma_{jk} = \omega_{mk}\sqrt{\omega_{mm}\omega_{kk}}$.

In this simulation, we used $n = 300$ samples and $p = 500$ TGs and considered two types of networks: hierarchical network and Barabasi-Albert network. For the hierarchical network, there were 5 disjointed subnetworks and each of them consisted of 100 TGs. The subnetwork was constructed as the same as Kim et al.[45] (**Figure D.1**). For Barabasi-Albert network, there were 50 subnetworks and each of them consisted of 10 TGs. For each subnetwork, a BA-based network was generated[46]. For both types of networks, the network structure $\mathbf{A} = (a_{mk})$ of 500 TGs was constructed. $a_{mk} = 1$ if m^{th} TG and k^{th} TG were within the same subnetwork and $a_{mk} = 0$ otherwise.

The regulation effects $\boldsymbol{\beta}$ were determined based on the relationship between TGs and the TF. In the hierarchical network, only 45 TGs in the first subnetwork, which contained one centered TG and four groups of TGs denoted as g_1, g_2, g_3, and g_4, were regulated by the TF. Therefore, the regulation effects $\boldsymbol{\beta}$ were defined as

$$\beta_j = \begin{cases} \beta, & if\ TG\ j\ is\ the\ centered\ TG, \\ \beta/3 \times \sqrt{d_j}, & if\ j \in g_1\ or\ j \in g_3, \\ -\beta/3 \times \sqrt{d_j}, & if\ j \in g_2\ or\ j \in g_4, \\ 0, & otherwis. \end{cases}$$

where d_j is the degree of TG j, which represents the number of TGs that were linked with TG j. In the Barabasi-Albert (BA)-based network, only 40 TGs in the first four subnetworks denoted as g_1, g_2, g_3, and g_4, were regulated by the TF. Therefore, the regulation effects $\boldsymbol{\beta}$ were defined as

$$\beta_j = \begin{cases} \beta \times \sqrt{d_j}, & if\ j \in g_1\ or\ j \in g_3, \\ -\beta \times \sqrt{d_j}, & if\ j \in g_2\ or\ j \in g_4, \\ 0, & otherwise. \end{cases}$$

Text D.3. APGD algorithm to solve Huber-Lasso.

In Huber-Lasso, we considered the Huber loss function and the Lasso penalty. Therefore, the penalized loss function can be decomposed as

$$f(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + h(\boldsymbol{\beta}) = \left(\sum_{i=1}^{n} H_M(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}) \right) + (\lambda \|\boldsymbol{\beta}\|_1). \qquad \text{(S1.1)}$$

where $g(\boldsymbol{\beta})$ and $h(\boldsymbol{\beta})$ are given by

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{n} H_M(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}), \qquad \text{(S1.2)}$$

$$h(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1. \tag{S1.3}$$

The APGD in k^{th} iteration can be defined as

$$\begin{aligned}\boldsymbol{\xi}^{k+1} &:= \boldsymbol{\beta}^k + \omega^k(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})\\ \boldsymbol{\theta}^{k+1} &:= \boldsymbol{\xi}^{k+1} - \gamma^k \nabla g(\boldsymbol{\xi}^{k+1})\\ \boldsymbol{\beta}^{k+1} &:= \mathbf{Prox}_{\gamma^k h}(\boldsymbol{\theta}^{k+1})\end{aligned} \tag{S1.4}$$

where $\omega^k \in [0,1)$ is an extrapolation parameter and γ^k is the usual step size. These parameters must be chosen in specific ways to achieve convergence acceleration. One simple choice [47] for ω^k is $k/(k+3)$. Here $\nabla g(\boldsymbol{\xi}^{k+1})$ is the gradient of the convex differentiable function $g(\cdot)$ at $\boldsymbol{\xi}^{k+1}$, which can be calculated by

$$\nabla g(\boldsymbol{\xi}^{k+1}) = \sum_{i=1}^{n} -\nabla H_M(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\xi}^{k+1})\boldsymbol{x}_i. \tag{S1.5}$$

where let $\Delta_i := y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\xi}^{k+1}$, then the gradient of Huber function can be calculated as $\nabla H_M(\Delta_i) = 2\Delta_i I(|\Delta_i| \le M) + 2M sign(\Delta_i) I(|\Delta_i| > M)$. The operator $\mathbf{Prox}_{\gamma^k h}(\boldsymbol{\theta}^{k+1})$ is called proximal mapping for $h(\boldsymbol{\beta})$. To solve the Huber-Lasso, the key is to compute the proximal mapping for the convex non-differentiable function $h(\boldsymbol{\beta})$. It is not difficult to verify[48]:

$$\begin{aligned}\mathbf{Prox}_{\gamma^k h}(\boldsymbol{\theta}^{k+1}) &= \text{argmin}_{\boldsymbol{\beta}}\left\{\lambda\|\boldsymbol{\beta}\|_1 + \frac{1}{2\gamma^k}\|\boldsymbol{\beta} - \boldsymbol{\theta}^{k+1}\|_2^2\right\}\\ &= \text{sign}(\boldsymbol{\theta}^{k+1})\max\{\|\boldsymbol{\theta}^{k+1}\|_1 - \gamma^k\lambda, 0\}.\end{aligned} \tag{S1.6}$$

To obtain a valid estimation in each iteration, we also defined an upper bound of $g(\cdot)$ as $\hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$ which is given by

$$\hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1}) = g(\boldsymbol{\xi}^{k+1}) + \nabla g(\boldsymbol{\xi}^{k+1})^T(\boldsymbol{\beta} - \boldsymbol{\xi}^{k+1}) + \frac{1}{2\gamma^k}\|\boldsymbol{\beta} - \boldsymbol{\xi}^{k+1}\|_2^2. \tag{S1.7}$$

Algorithm S1 APGD for Huber-Lasso

function APGD.HuberLasso($\mathbf{X}, \boldsymbol{y}, \lambda$)
  Initiate $\boldsymbol{\beta}^0 = \boldsymbol{\beta}^1 = \mathbf{0}, \gamma = 1000$
  for $k \in 1 \ldots$ niter **do**
    $\boldsymbol{\xi}^{k+1} \leftarrow \boldsymbol{\beta}^k + k/(k+3) \times (\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$
    Compute $\nabla g(\boldsymbol{\xi}^{k+1})$ from (S1.5)
    while TRUE **do**
      Compute $\boldsymbol{\beta}^{prox}$ from (S1.4)
      Compute $g(\boldsymbol{\beta}^{prox})$ from (S1.2) and $\hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$ from (S1.7)
      break if $g(\boldsymbol{\beta}^{prox}) \le \hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$
      $\gamma \leftarrow \gamma \times 0.5$
    $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\beta}^{prox}$
    break if $|f(\boldsymbol{\beta}^{prox}) - f(\boldsymbol{\beta}^k)| < 10^{-8}$ or $|\boldsymbol{\beta}^{prox} - \boldsymbol{\beta}^k| < 10^{-4}$
  return $\boldsymbol{\beta}^{prox}$

Text D.4. APGD algorithm to solve Huber-ENET.

In Huber-ENET, we considered the Huber loss function and the Elastic Net penalty. Therefore, the convex differentiable function $g(\boldsymbol{\beta})$ and the convex non-differentiable function $h(\boldsymbol{\beta})$ are given by

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{n} H_M(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta}) + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\boldsymbol{\beta}, \qquad h(\boldsymbol{\beta}) = \lambda\alpha\|\boldsymbol{\beta}\|_1. \quad \text{(S2.1)}$$

Therefore, the proximal operator in APGD for Huber-ENET and the gradient of convex differentiable function $g(\cdot)$ at $\boldsymbol{\xi}^{k+1}$, which can be calculated by the following formulas.

$$\mathbf{Prox}_{\gamma^k h}(\boldsymbol{\theta}^{k+1}) = \text{sign}(\boldsymbol{\theta}^{k+1})\max\{\|\boldsymbol{\theta}^{k+1}\|_1 - \gamma^k\lambda\alpha, 0\} \quad \text{(S2.2)}$$

$$\nabla g(\boldsymbol{\xi}^{k+1}) = \sum_{i=1}^{n} -\nabla H_M(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\xi}^{k+1})\boldsymbol{x}_i + \lambda(1-\alpha)\boldsymbol{\xi}^{k+1} \quad \text{(S2.3)}$$

Algorithm S2 APGD for Huber-ENET

function APGD.HUBERENET($\mathbf{X}, \boldsymbol{y}, \lambda, \alpha$)
  Initiate $\boldsymbol{\beta}^0 = \boldsymbol{\beta}^1 = \mathbf{0}, \gamma = 1000$
  for $k \in 1 \ldots$ niter **do**
    $\boldsymbol{\xi}^{k+1} \leftarrow \boldsymbol{\beta}^k + k/(k+3) \times (\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$
    Compute $\nabla g(\boldsymbol{\xi}^{k+1})$ from (S2.3)
    while TRUE **do**
      Compute $\boldsymbol{\beta}^{prox}$ from (S2.2)
      Compute $g(\boldsymbol{\beta}^{prox})$ from (S2.1) and $\hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$ from (S1.7)
      break if $g(\boldsymbol{\beta}^{prox}) \leq \hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$
      $\gamma \leftarrow \gamma \times 0.5$
    $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\beta}^{prox}$
    break if $|f(\boldsymbol{\beta}^{prox}) - f(\boldsymbol{\beta}^k)| < 10^{-8}$ or $|\boldsymbol{\beta}^{prox} - \boldsymbol{\beta}^k| < 10^{-4}$
  return $\boldsymbol{\beta}^{prox}$

Text D.5. APGD algorithm to solve Huber-Net.

In Huber-Net, we considered the Huber loss function and the network-based penalty. Therefore, the convex differentiable function $g(\boldsymbol{\beta})$ and the convex non-differentiable function $h(\boldsymbol{\beta})$ are given by

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{n} H_M(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta}) + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\mathbf{S}^T\mathbf{L}\mathbf{S}\boldsymbol{\beta}, \quad h(\boldsymbol{\beta}) = \lambda\alpha\|\boldsymbol{\beta}\|_1. \quad \text{(S3.1)}$$

Therefore, the proximal operator in APGD for Huber-Net and the gradient of convex differentiable function $g(\cdot)$ at $\boldsymbol{\xi}^{k+1}$, which can be calculated by the following formulas.

$$\mathbf{Prox}_{\gamma^k h}(\boldsymbol{\theta}^{k+1}) = \text{sign}(\boldsymbol{\theta}^{k+1})\max\{\|\boldsymbol{\theta}^{+1}\|_1 - \gamma^k\lambda\alpha, 0\} \quad \text{(S3.2)}$$

$$\nabla g(\boldsymbol{\xi}^{k+1}) = \sum_{i=1}^{n} -\nabla H_M(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\xi}^{k+1})\boldsymbol{x}_i + \lambda(1-\alpha)\mathbf{S}^T\mathbf{L}\mathbf{S}\boldsymbol{\xi}^{\boldsymbol{k+1}}. \quad \text{(S3.3)}$$

Algorithm S3 APGD for Huber-Net

function APGD.HUBERNET($\mathbf{X}, \boldsymbol{y}, \mathbf{L}, \mathbf{S}, \lambda, \alpha$)
  Initiate $\boldsymbol{\beta}^0 = \boldsymbol{\beta}^1 = \mathbf{0}, \gamma = 1000$
  for $k \in 1 \ldots$ niter **do**
    $\boldsymbol{\xi}^{k+1} \leftarrow \boldsymbol{\beta}^k + k/(k+3) \times (\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$
    Compute $\nabla g(\boldsymbol{\xi}^{k+1})$ from (S3.3)
    while TRUE **do**
      Compute $\boldsymbol{\beta}^{prox}$ from (S3.2)
      Compute $g(\boldsymbol{\beta}^{prox})$ from (S3.1) and $\hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$ from (S1.7)
      break if $g(\boldsymbol{\beta}^{prox}) \leq \hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$
      $\gamma \leftarrow \gamma \times 0.5$
    $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\beta}^{prox}$
    break if $|f(\boldsymbol{\beta}^{prox}) - f(\boldsymbol{\beta}^k)| < 10^{-8}$ or $|\boldsymbol{\beta}^{prox} - \boldsymbol{\beta}^k| < 10^{-4}$
  return $\boldsymbol{\beta}^{prox}$

Text D.6. APGD algorithm to solve MSE-Lasso.

In MSE-Lasso, we considered the MSE loss function and the Lasso penalty. Therefore, the convex differentiable function $g(\boldsymbol{\beta})$ and the convex non-differentiable function $h(\boldsymbol{\beta})$ are given by

$$g(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})^2, \qquad h(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1. \quad \text{(S4.1)}$$

Therefore, the proximal operator in APGD for MSE-Lasso and the gradient of convex differentiable function $g(\cdot)$ at $\boldsymbol{\xi}^{k+1}$, which can be calculated by the following formulas.

$$\mathbf{Prox}_{\gamma^k h}(\boldsymbol{\theta}^{k+1}) = \text{sign}(\boldsymbol{\theta}^{k+1})\max\{\|\boldsymbol{\theta}^{k+1}\|_1 - \gamma^k\lambda, 0\} \quad \text{(S4.2)}$$

$$\nabla g(\boldsymbol{\xi}^{k+1}) = \frac{1}{n}\sum_{i=1}^{n} -(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\xi}^{k+1})\boldsymbol{x}_i. \quad \text{(S4.3)}$$

Algorithm S4 APGD for MSE-Lasso

function APGD.MSELASSO($\mathbf{X}, \boldsymbol{y}, \lambda$)
  Initiate $\boldsymbol{\beta}^0 = \boldsymbol{\beta}^1 = \mathbf{0}, \gamma = 1000$
  for $k \in 1 \ldots$ niter **do**
    $\boldsymbol{\xi}^{k+1} \leftarrow \boldsymbol{\beta}^k + k/(k+3) \times (\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$
    Compute $\nabla g(\boldsymbol{\xi}^{k+1})$ from (S4.3)
    while TRUE **do**
      Compute $\boldsymbol{\beta}^{prox}$ from (S4.2)
      Compute $g(\boldsymbol{\beta}^{prox})$ from (S4.1) and $\hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$ from (S1.7)
      break if $g(\boldsymbol{\beta}^{prox}) \leq \hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$
      $\gamma \leftarrow \gamma \times 0.5$
    $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\beta}^{prox}$
    break if $|f(\boldsymbol{\beta}^{prox}) - f(\boldsymbol{\beta}^k)| < 10^{-8}$ or $|\boldsymbol{\beta}^{prox} - \boldsymbol{\beta}^k| < 10^{-4}$
  return $\boldsymbol{\beta}^{prox}$

Text D.7. APGD algorithm to solve MSE-ENET.

In MSE-ENET, we considered the MSE loss function and the Elastic Net penalty. Therefore, the convex differentiable function $g(\boldsymbol{\beta})$ and the convex non-differentiable function $h(\boldsymbol{\beta})$ are given by

$$g(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\boldsymbol{\beta}, \qquad h(\boldsymbol{\beta}) = \lambda\alpha\|\boldsymbol{\beta}\|_1. \quad \text{(S5.1)}$$

Therefore, the proximal operator in APGD for MSE-ENET and the gradient of convex differentiable function $g(\cdot)$ at $\boldsymbol{\xi}^{k+1}$, which can be calculated by the following formulas.

$$\mathbf{Prox}_{\gamma^k h}(\boldsymbol{\theta}^{k+1}) = \text{sign}(\boldsymbol{\theta}^{k+1})\max\{\|\boldsymbol{\theta}^{k+1}\|_1 - \gamma^k\lambda\alpha, 0\} \quad \text{(S5.2)}$$

$$\nabla g(\boldsymbol{\xi}^{k+1}) = \frac{1}{n}\sum_{i=1}^{n} -(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\xi}^{k+1})\boldsymbol{x}_i + \lambda(1-\alpha)\boldsymbol{\xi}^{k+1} \quad \text{(S5.3)}$$

Algorithm S5 APGD for MSE-ENET

function APGD.MSEENET($\mathbf{X}, \boldsymbol{y}, \lambda, \alpha$)
  Initiate $\boldsymbol{\beta}^0 = \boldsymbol{\beta}^1 = \mathbf{0}, \gamma = 1000$
  for $k \in 1 \ldots$ niter **do**
    $\boldsymbol{\xi}^{k+1} \leftarrow \boldsymbol{\beta}^k + k/(k+3) \times (\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$
    Compute $\nabla g(\boldsymbol{\xi}^{k+1})$ from (S5.3)
    while TRUE **do**
      Compute $\boldsymbol{\beta}^{prox}$ from (S5.2)
      Compute $g(\boldsymbol{\beta}^{prox})$ from (S5.1) and $\hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$ from (S1.7)
      break if $g(\boldsymbol{\beta}^{prox}) \leq \hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$
      $\gamma \leftarrow \gamma \times 0.5$
    $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\beta}^{prox}$
    break if $|f(\boldsymbol{\beta}^{prox}) - f(\boldsymbol{\beta}^k)| < 10^{-8}$ or $|\boldsymbol{\beta}^{prox} - \boldsymbol{\beta}^k| < 10^{-4}$
  return $\boldsymbol{\beta}^{prox}$

Text D.8. APGD algorithm to solve MSE-Net.

In MSE-Net, we considered the MSE loss function and the network-based pnalty. Therefore, the convex differentiable function $g(\boldsymbol{\beta})$ and the convex non-differentiable function $h(\boldsymbol{\beta})$ are given by

$$g(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\mathbf{S}^T\mathbf{L}\mathbf{S}\boldsymbol{\beta}, \quad \text{(S6.1)}$$
$$h(\boldsymbol{\beta}) = \lambda\alpha\|\boldsymbol{\beta}\|_1.$$

Therefore, the proximal operator in APGD for Huber-Net and the gradient of convex differentiable function $g(\cdot)$ at $\boldsymbol{\xi}^{k+1}$, which can be calculated by the following formulas.

$$\mathbf{Prox}_{\gamma^k h}(\boldsymbol{\theta}^{k+1}) = \text{sign}(\boldsymbol{\theta}^{k+1})\max\{\|\boldsymbol{\theta}^{k+1}\|_1 - \gamma^k\lambda\alpha, 0\} \quad \text{(S6.2)}$$

$$\nabla g(\boldsymbol{\xi}^{k+1}) = \frac{1}{n}\sum_{i=1}^{n} -(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\xi}^{k+1})\boldsymbol{x}_i + \lambda(1-\alpha)\mathbf{S}^T\mathbf{L}\mathbf{S}\boldsymbol{\xi}^{\boldsymbol{k+1}}. \quad \text{(S6.3)}$$

Algorithm S6 APGD for MSE-Net

function APGD.MSENET($\mathbf{X}, \boldsymbol{y}, \mathbf{L}, \mathbf{S}, \lambda, \alpha$)
  Initiate $\boldsymbol{\beta}^0 = \boldsymbol{\beta}^1 = \mathbf{0}, \gamma = 1000$
  for $k \in 1 \ldots$ niter **do**
    $\boldsymbol{\xi}^{k+1} \leftarrow \boldsymbol{\beta}^k + k/(k+3) \times (\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1})$
    Compute $\nabla g(\boldsymbol{\xi}^{k+1})$ from (S6.3)
    while TRUE **do**
      Compute $\boldsymbol{\beta}^{prox}$ from (S6.2)
      Compute $g(\boldsymbol{\beta}^{prox})$ from (S6.1) and $\hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$ from (S1.7)
      break if $g(\boldsymbol{\beta}^{prox}) \leq \hat{g}_{\gamma^k}(\boldsymbol{\beta}, \boldsymbol{\xi}^{k+1})$
      $\gamma \leftarrow \gamma \times 0.5$
    $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\beta}^{prox}$
    break if $|f(\boldsymbol{\beta}^{prox}) - f(\boldsymbol{\beta}^k)| < 10^{-8}$ or $|\boldsymbol{\beta}^{prox} - \boldsymbol{\beta}^k| < 10^{-4}$
  return $\boldsymbol{\beta}^{prox}$

Text D.9. Implementation of APGD and TGPred.

Six statistical selection methods based on the penalized regression models and the APGD algorithm for solving these six statistical methods had been implemented in both Python3 and R and then packed into TGPred. Both of them used commonly used libraries for scientific computing. For Python3 version of TGPred, we used numpy, scipy, and sklearn to support efficient mathematical and dataframe computing, cvxpy to compare the runtime and estimated results of APGD with commonly used CVX, and networkx to generate synthetic data based on the BA network setting. For R version of TGPred, we used Matrix and MASS to support the efficient mathematical computing, and mvtnorm and igraph to generate synthetic data. TGPred can be directly used within Python and R. Both regulation effect β_j and selection probability SP_j of target gene j can be calculated by TGPred for $j = 1, \cdots, p$. Note that the large-scale genetic data set is acceptable to APGD and the computation time was evaluated on the high-performance computing (HPC) cluster (Intel Xeon E5-2670 2.6 GHz, 16 GB RAM). For example, when the number of TGs are greater than 30,000 ($p > 30{,}000$) and $B = 500$ times of half-sample approach, the computation times of ENET penalty along with MSE and Huber loss functions for all genes were about 12h CPU time with 90 pairs of tuning parameters α and λ; the computation times of Lasso penalty was about 8h CPU time with 50 tuning parameters λ; and the computation times of Net penalty was about 26h CPU time with 90 pairs of tuning parameters α and λ. TGPred packages have been made publicly available on GitHub as open-source software for downloading (https://github.com/xueweic/TGPred); more detailed information on how to install and run the tool was enclosed in the packages.

Text D.10. Comparison of computational time and regression coefficients estimated by APGD and CVX.

We also compared the computational efficiency and the regression coefficients estimated by APGD and CVX, a commonly used package for convex optimization, for several pairs of tuning parameters λ and α. **Figures D.4-D.6** showed that the computation times of CVX and APGD among all grid sets of α and λ based on $B = 500$ subsamples drawn with the half-sample approach. **Figure D.4** showed the computation times of Huber-Lasso, Huber-ENET, MSE-Lasso, and MSE-ENET under the general setting with $\beta = 0.2$. For ENET penalty function, $n_\lambda = 1, \cdots, 10$ indicated the order of selected λ in a log10-scale from $ratio * \lambda_{max}$ to λ_{max}, where λ_{max} is related to $\alpha = 0.1, \cdots, 0.9$. For Lasso penalty, $n_\lambda = 1, \cdots, 100$ indicated the order of selected λ in a log10-scale from $ratio * \lambda_{max}$ to λ_{max}, where λ_{max} is related to $\alpha = 1$. The datasets were simulated under the same setting (**Text D.1**). All analyses were performed on a macOS (2.7 GHz Quad-Core Intel Core i7, 16 GB memory). It can be seen that APGD is much more computationally efficient than CVX since the running time of APGD was less than one fifth time of CVX for all six methods (**Figure D.4**). A disadvantage of CVX is that all of the estimated regression coefficients are not equal to 0 (around 10^{-22} for non-zero regression coefficients). Therefore, the stability selection method may not be applicable to the CVX method since it is difficult to find a cut-off threshold for the regression coefficients. The APGD algorithm was also evaluated under the hierarchical network and Barabasi-Albert network settings. As shown in **Figures D.5-D.6**, the computation times of APGD were much shorter than those of CVX no matter which methods (Huber-Lasso, Huber-ENET, Huber-Net, MSE-Lasso, MSE-ENET, and MSE-Net) it was applied to. The results manifested that APGD was consistently more computational efficient than CVX, as we had observed for the general setting.

We compared the regression coefficients estimated by APGD and CVX for several pairs of tuning parameters λ and α. **Figures D.7-D.9** showed that the QQ plots of the regression coefficients estimated by both CVX and APGD. **Figure D.7** showed the estimation of regulation effects of Huber-Lasso, Huber-ENET, MSE-Lasso, and MSE-ENET under the general setting with $\beta = 0.2$. The values lied along the diagonal line as the Huber loss function was used, indicating the regression coefficients estimated by CVX and APGD were identical. When the MSE loss function was used, the non-zero estimations of regulation effects of CVX were greater than that of APGD (**Fig D.7**). However, there were only 50 true TGs (out of 500 genes) that were regulated by a given TF in this simulation setting. That is, CVX obtained more false positives than APGD. Except for those false positives estimated by CVX, the regression coefficients estimated by these two methods were nearly identical. **Figures D.8-D.9** showed that the estimation of regulation effects of our proposed six statistical selection methods under the network setting, where we used $\beta = 0.4$ in the hierarchical network setting (**Figure D.8**) and $\beta = 0.1$ in the Barabasi-Albert network setting (**Figure D.9**). We observed that the patterns of the estimation performance were similar to that shown in **Figure D.7**.

D.2 Supplementary Figures

Figure D.1. The hierarchical network module is used in the hierarchical network setting. There is a total of 100 genes that contain a centered gene.

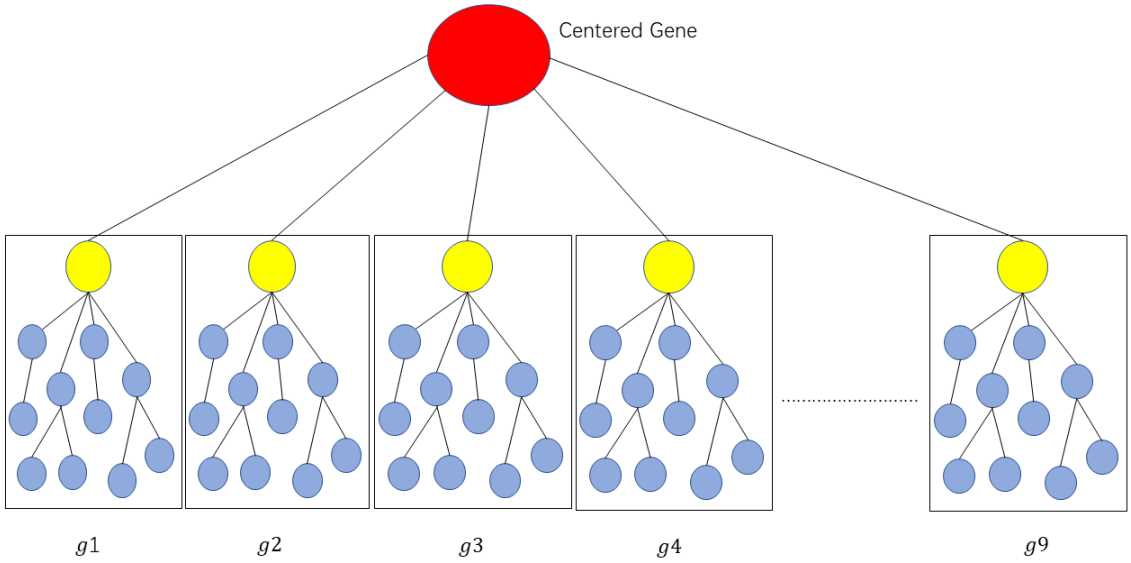


Figure D.2. The AuROC of the selection probabilities of the different methods in the general setting, which corresponding to Figure 4.1.

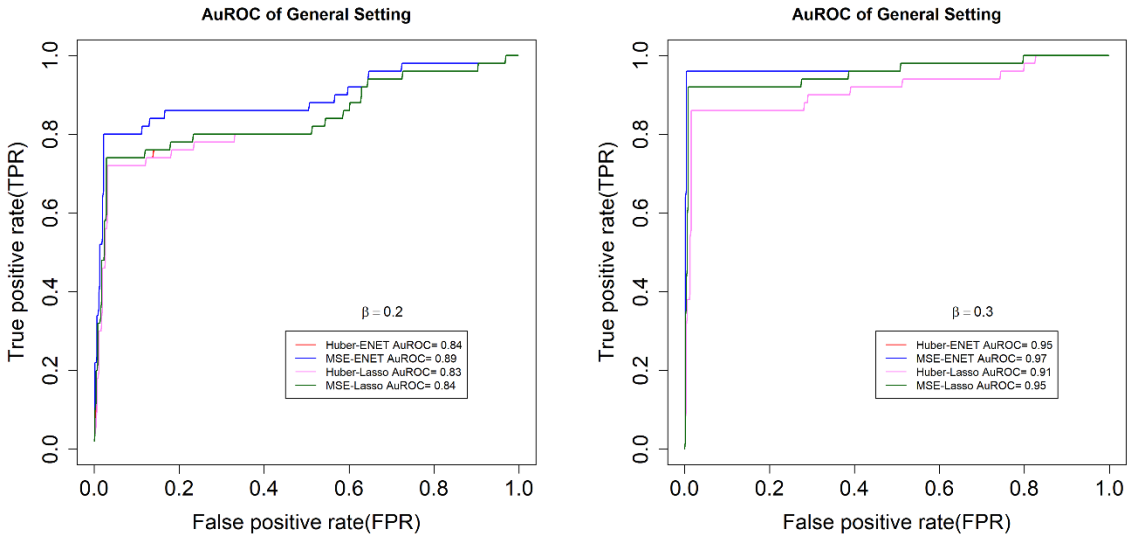


Figure D.3. The AuROC of the selection probabilities of the different methods in the network setting, which corresponding to Figure 4.2.

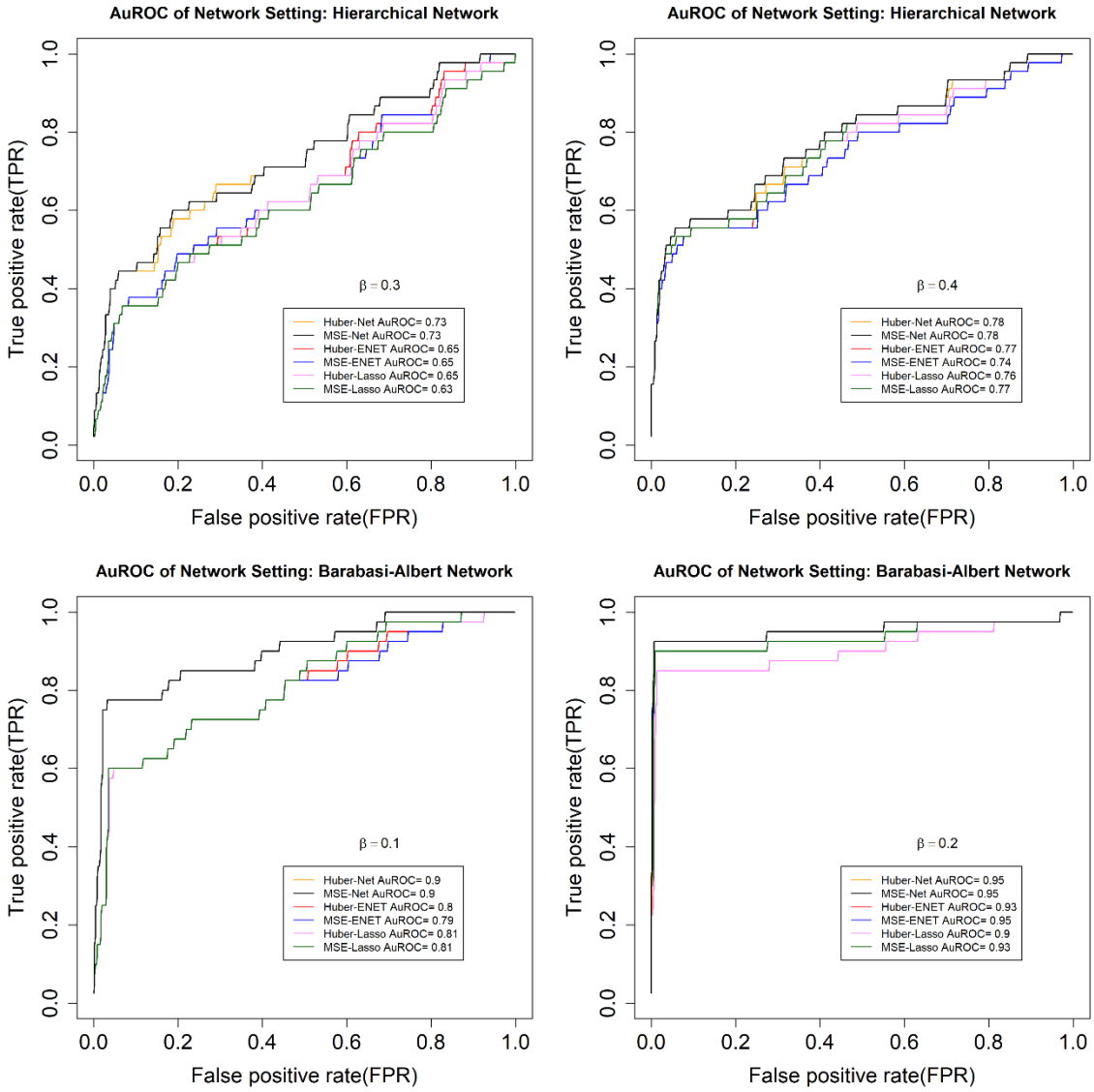


Figure D.4. The computation times of CVX versus APGD in the general setting ($\beta = 0.2$) among all grid sets of α and λ based on half-sample approach with $B = 500$ times of resampling.

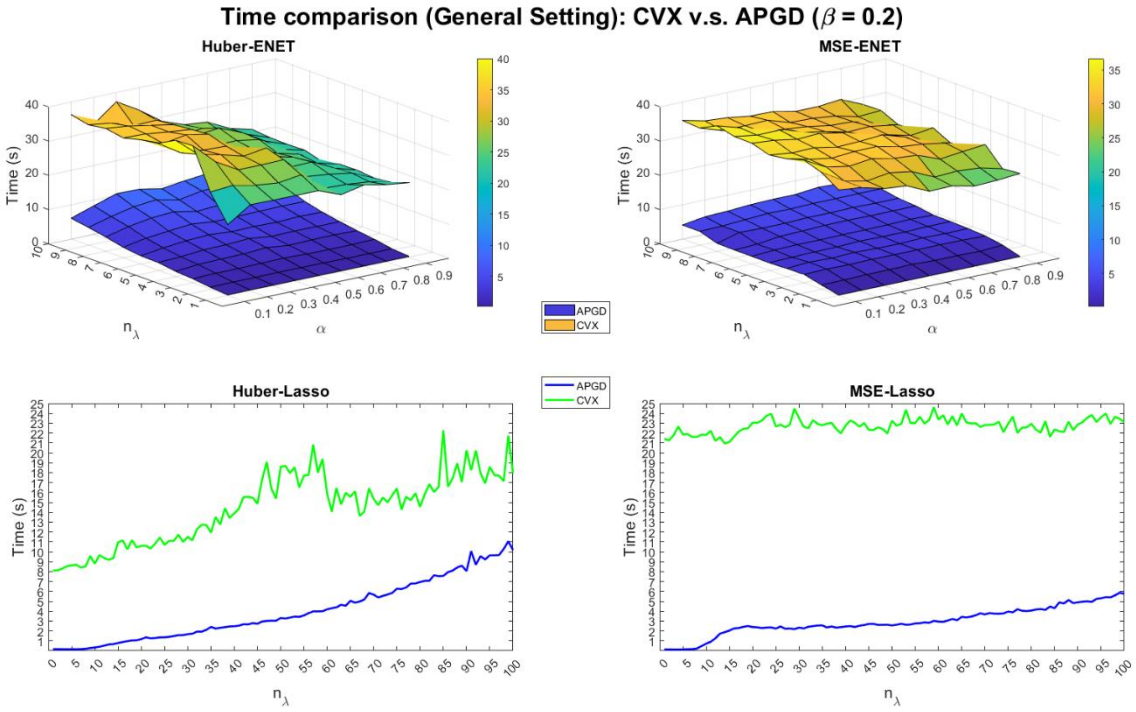


Figure D.5. The computation times of CVX versus APGD in the hierarchical network setting ($\beta = 0.4$) among all grid sets of α and λ based on half-sample approach with $B = 500$ times of resampling.

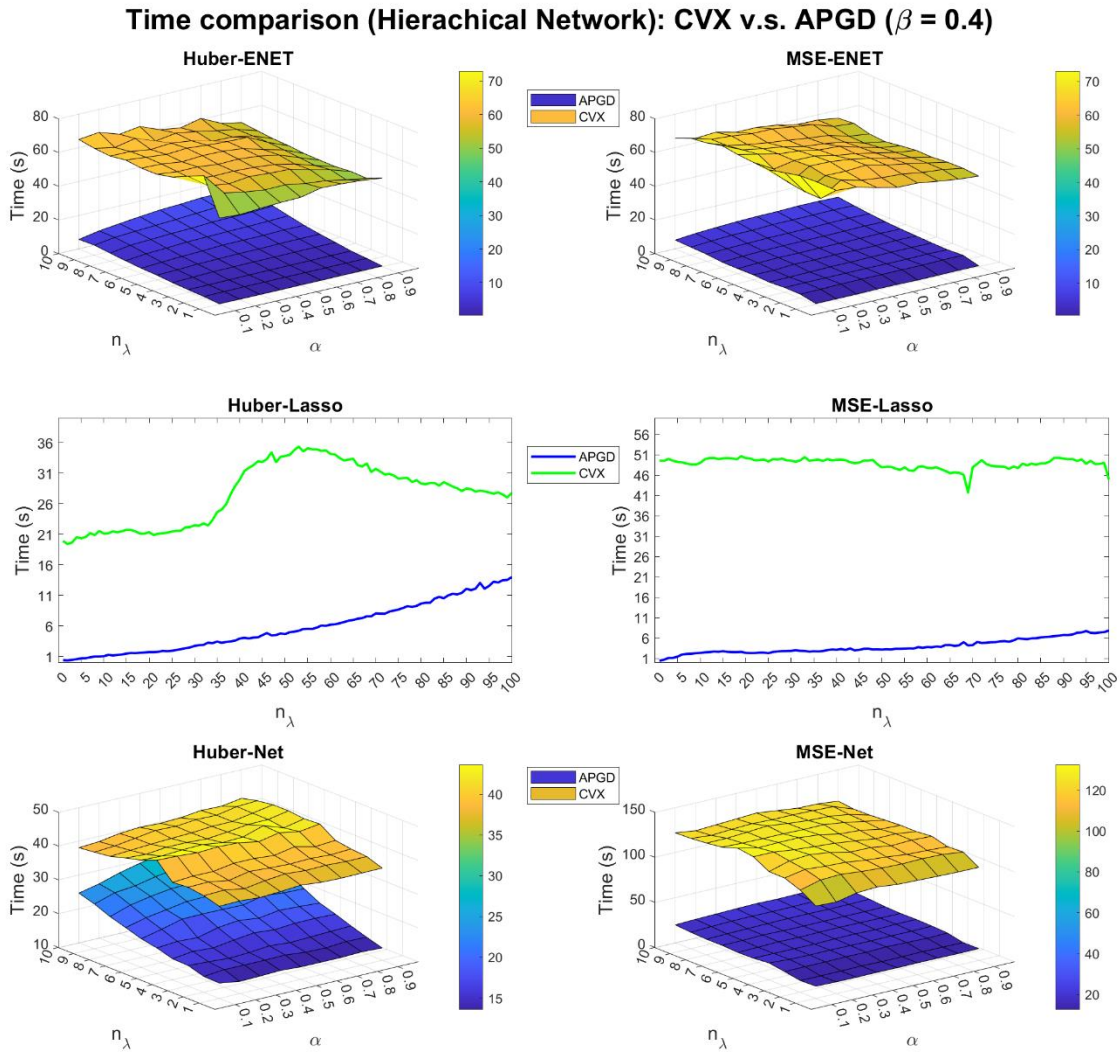


Figure D.6. The computation times of CVX versus APGD in the Barabasi-Albert network setting ($\beta = 0.1$) among all grid sets of α and λ based on half-sample approach with $B = 500$ times of resampling.

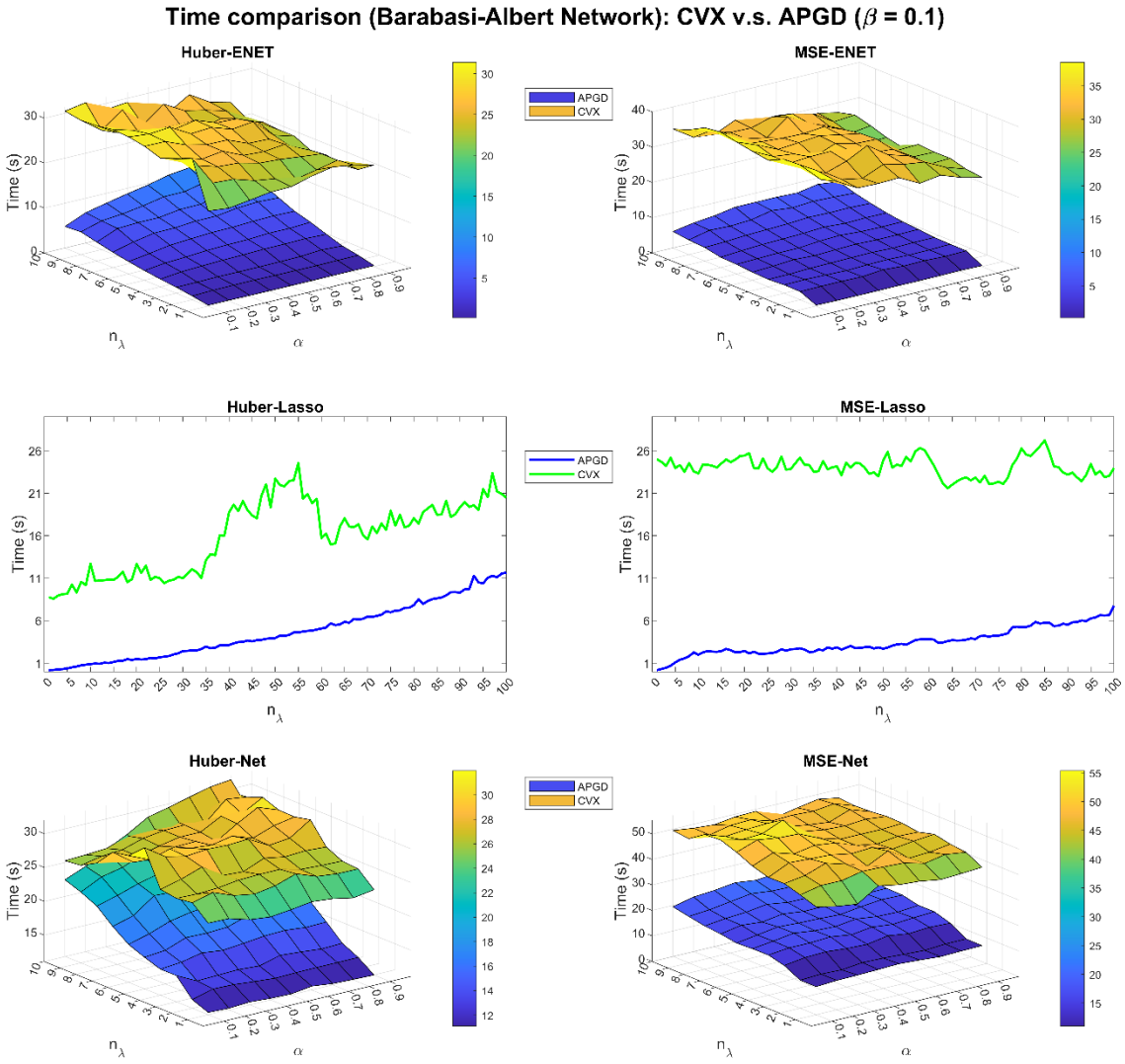


Figure D.7. The estimation of regulation effects (beta) comparison of CVX versus APGD in the general setting ($\beta = 0.2$) by different algorithms.

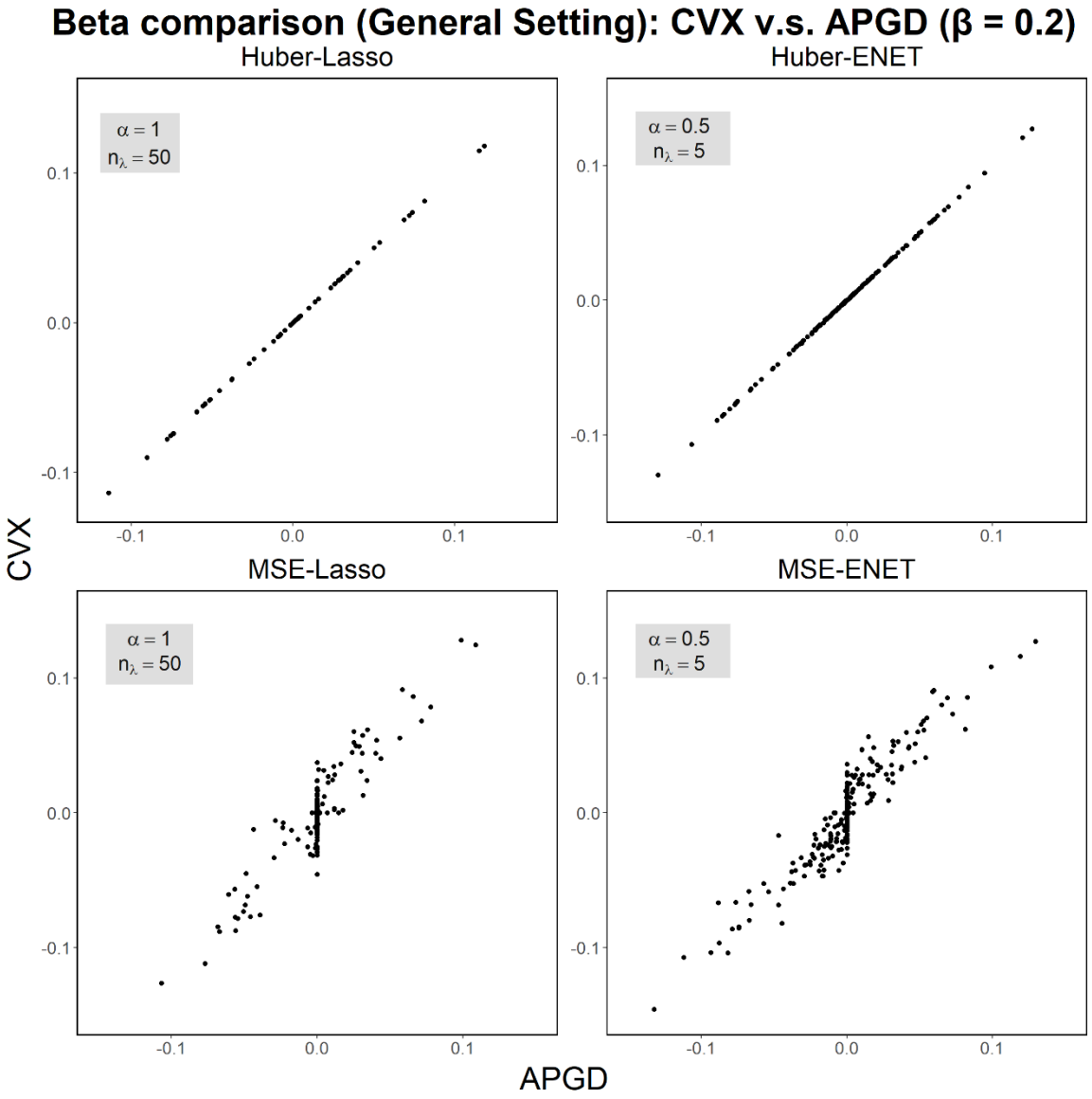


Figure D.8. The estimation of regulation effects (beta) comparison of CVX versus APGD in the hierarchical network setting ($\beta = 0.4$) by different algorithms.

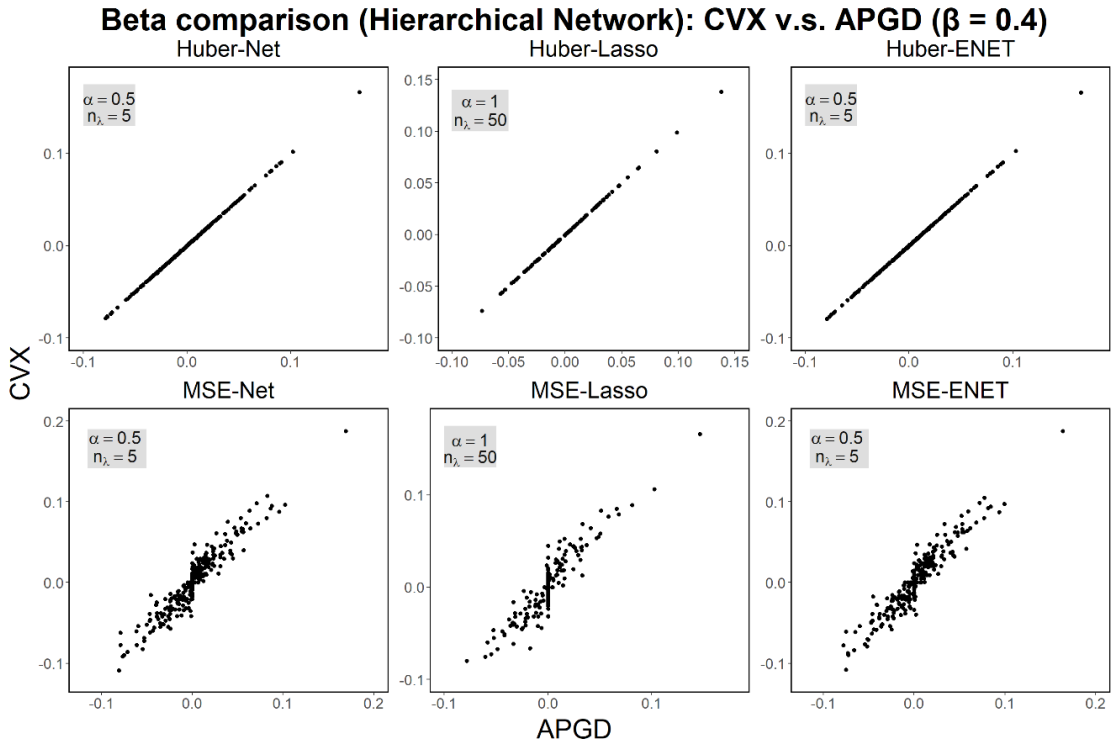


Figure D.9. The estimation of regulation effects (beta) comparison of CVX versus APGD in the Barabasi-Albert network setting ($\beta = 0.1$) by different algorithms.

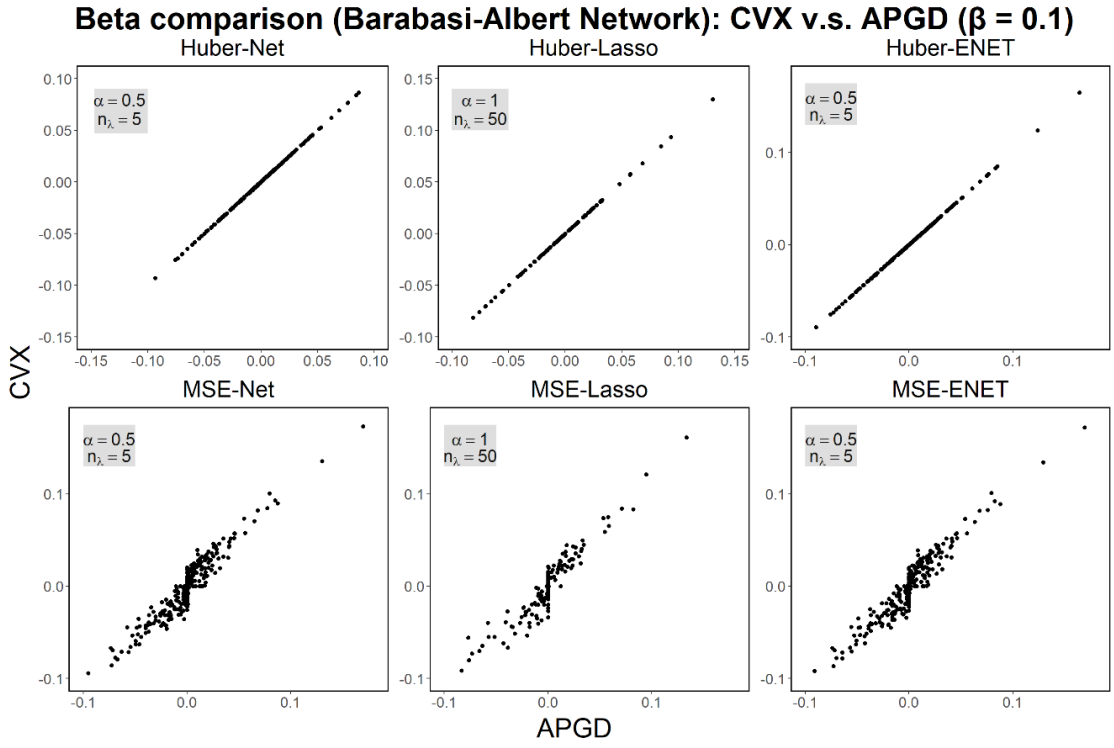


Figure D.10. Venn diagram representing the numbers of common and unique target genes of each of the 23 TFs identified by the three methods with Huber loss function (Huber-ENET, Huber-Lasso and Huber-Net) and three methods with MSE loss function (MSE-ENET, MSE-Lasso, and MSE-Net).

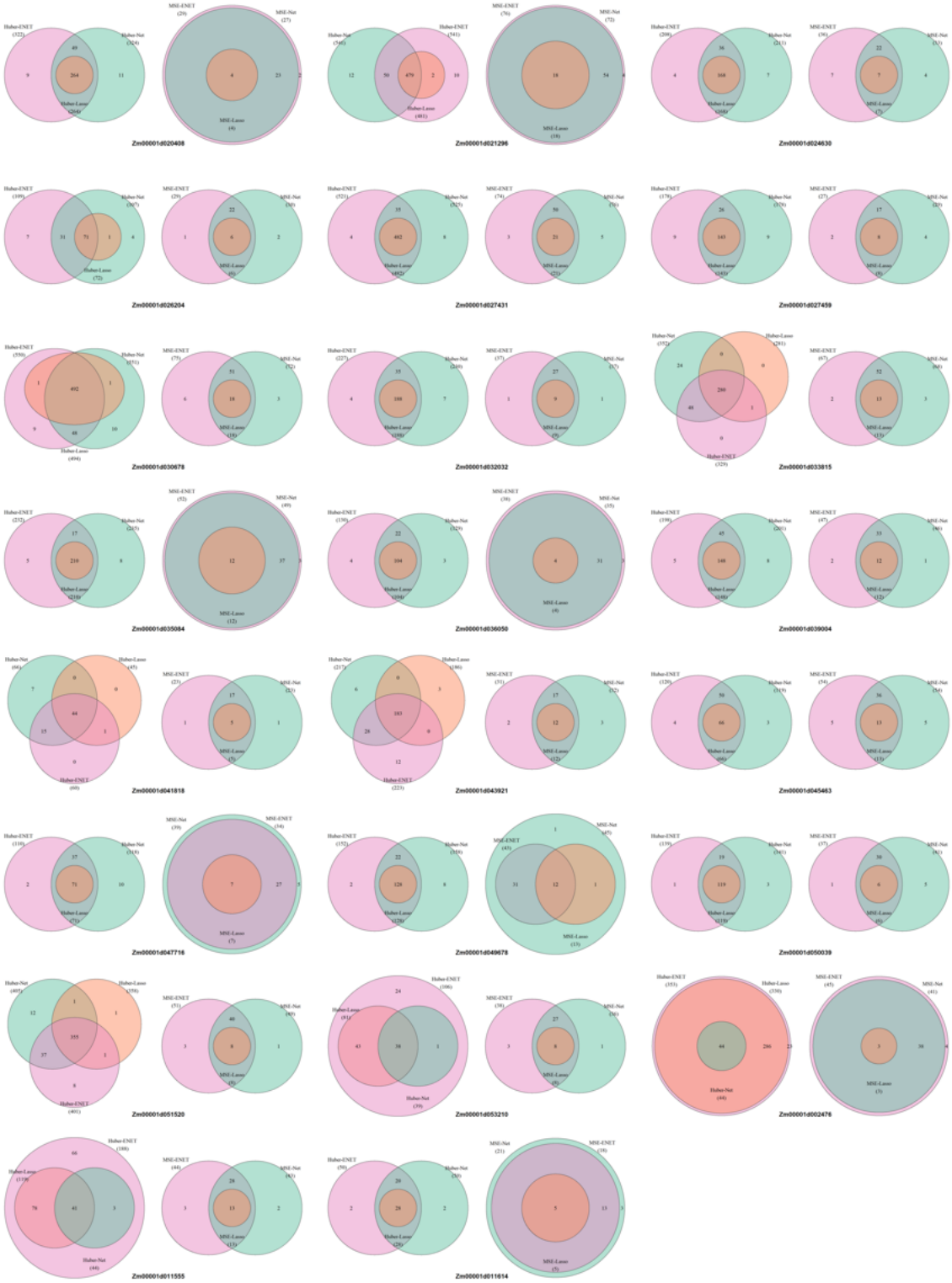
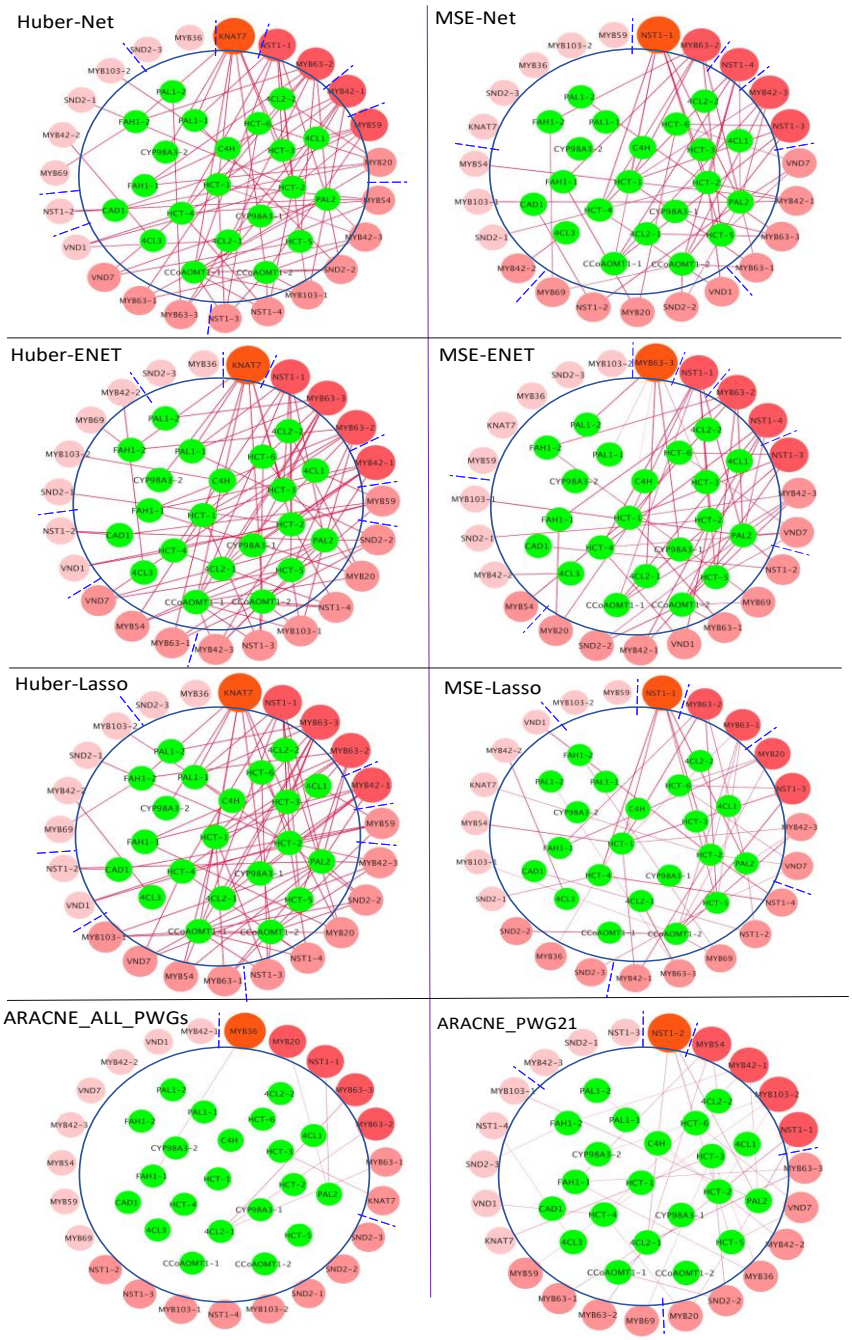


Figure D.11. The gene regulatory networks of lignin pathway genes produced by the six methods, Huber-ENET, Huber-Lasso, Huber-Net, MSE-ENET, MSE-Lasso, MSE-Net, where regulatory genes were ranked based on their connectivities to pathway genes in clockwise. The inputs were the expression data sets of 2539 pathway genes (PWGs) and 23 known lignin pathway regulators in the in maize. The network of ARACNE_ALL_PWGs was produced by ARANCE method with the same inputs as the six methods we developed, while the network of ARACNE_PWG21was produced by ARANCE method with the expression data sets of 21 lignin pathway genes and 23 known lignin pathway regulators being used as the inputs.



E Supplementary Materials for Chapter 5

E.1 Supplementary Texts

Text E.1. Simulation Setups.

The individual-level genetic data where linked genes within a biological network are correlated with each other are generated using the following three steps:

Step 1: Construct an M dimensional covariance matrix from an arbitrary graph based on a Gaussian graphical model.

Consider a total number of $M=1000$ genes that contain 10 disjointed network modules, each of which consists of 100 genes. Similar to Kim et al.[45], we construct each network module from **Figure E.1**, which contains a centered gene correlated with other genes with a few links in one network module. Therefore, the adjacency matrix $\mathbf{A}=(a_{mk})$ of those 1000 genes is constructed based on the connections among genes in each network module, where $a_{mk}=1$ represents the m^{th} and k^{th} genes are within the same network module and $a_{mk}=0$ otherwise. Next, we apply a Gaussian graphical model to generate a covariance matrix of 1000 genes[43]. Following the settings in Peng et al.[43], the initial concentration matrix $\mathbf{\Omega}=(\rho_{mk})_{M\times M}$ is generated by

$$\rho_{mk}=\begin{cases}1, & \text{if } m=k;\\ \sim U(D), & \text{if } m \text{ and } k \text{ are linked to each other;}\\ 0, & \text{otherwise,}\end{cases}$$

where $D=[-0.7,-0.1]\cup[0.1,0.7]$ and $U(D)$ represents a random variable from a uniform distribution on the domain D. We then rescale the non-zero elements in the concentration matrix to assure positive definiteness, that is, we divide each off-diagonal element by 1.5-fold of the sum. Finally, we average the rescaled matrix with its transpose to ensure symmetry and set the diagonal entries to be one. We denote the final matrix as $\mathbf{\Omega}$ and the covariance matrix $\mathbf{\Sigma}$ can be determined by $\Sigma_{mk}=\Omega^{-1}_{mk}\sqrt{\Omega^{-1}_{mm}\Omega^{-1}_{kk}}$, where Ω^{-1}_{mk} represents the $(m,k)^{th}$ element of the inversed concentration matrix $\mathbf{\Omega}^{-1}$. Note that the correlations between linked genes are much higher than that of unlinked genes.

Step 2: Generate M gene-level signals from different multivariate normal distributions for cases and controls, respectively.

In this step, we consider two scenarios to set up the phenotype-related genes. In scenario 1, we assume that only 45 genes in the first network module are phenotype related, where these 45 genes contain the centered gene and four subgroups of genes denoted as g_1, g_2, g_3, and g_4, respectively. In scenario 2, we assume that 48 genes in the first four network modules are phenotype related, where each of network modules contains one centered gene and a subgroup of genes which are denoted as g_1, g_2, g_3, and g_4,

respectively. For each scenario, let $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_M)^T$ be the mean vector, where $\boldsymbol{\mu} = (0, \cdots, 0)^T$ for the control group. In the case group, we set $\mu_m = 0$ for neutral genes (i.e., 955 genes in scenario 1 and 952 genes in scenario 2). In contrast, the mean of phenotype related genes μ_m is defined as

$$\mu_m = \begin{cases} \delta, & \text{if centered gene;} \\ \delta/3 \times \sqrt{d_m}, & \text{if } m \in g_1 \text{ or } g_3; \\ -\delta/3 \times \sqrt{d_m}, & \text{if } m \in g_2 \text{ or } g_4, \end{cases}$$

where δ is the strength of association signals and d_m is the total number of genetic links for the m^{th} gene. Therefore, the gene-level signals for each individual can be generated from a multivariate normal distribution, $\boldsymbol{z}_i \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i = 1, \cdots, n$.

Step 3: Generate DNA methylation and DNA sequence data based on each gene-level signal.

Consider $k_m = 10$ genetic components for $m = 1, \cdots, M$ and a total of 10,000 genetic components in simulation studies. In this step, we consider two types of genetic data, DNA methylation data and DNA sequence data. Let ω be the number of components correlated with the gene-level signal value z_{im} for the i^{th} individual and the m^{th} gene, which controls the number of causal or neutral components. The methylation value of the i^{th} individual and j^{th} CpG site in the m^{th} gene is denoted by x_{ij}^m which can be generated by

$$x_{ij}^m = \begin{cases} z_{im} + \varepsilon_{ij}, & j = 1, \cdots, \omega; \\ \bar{\varepsilon}_{ij}, & j = \omega + 1, \cdots, k_m, \end{cases}$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$ indicates the difference between the j^{th} CpG site and gene-level signal z_{im} and σ^2 is the error variance that controls the noise level of association signals. $\bar{\varepsilon}_{ij}$ follows a normal distribution with mean $\sum_{i=1}^{n} z_{im} / n$ and variance σ^2.

The value of genotype data usually indicates the genotypic score of an individual at a SNP which is the number of minor alleles that the individual carries at a SNP. We first generate two different continuous data $g_{ij,k}^m$ for $k = 1$ or 2 to indicate the genotypic value for two alleles which are defined as

$$g_{ij,k}^m = \begin{cases} z_{im} + \varepsilon_{ij}, & j = 1, \cdots, \omega; \\ \bar{\varepsilon}_{ij}, & j = \omega + 1, \cdots, J. \end{cases}$$

Next, we convert continuous data $g_{ij,k}^m$ to binary data $x_{ij,k}^m$ based on a fixed MAF_j^m for the j^{th} SNP in the m^{th} gene. Finally, the genotype data x_{ij}^m for the j^{th} SNP in the m^{th} gene is generated by $x_{ij}^m = x_{ij,1}^m + x_{ij,2}^m$. Therefore, the genotype data is coded as 0, 1, or 2. In our simulation studies, we consider five rare variants and five common variants in each gene, where the MAFs for rare variants are randomly generated from a uniform distribution $U(0.001, 0.01)$ and MAFs for common variants are from $U(0.01, 0.5)$.

Text E.2. Comparison of the Methods without Considering Network Structure.

After using the three weighted combinations (OWS, LD-PRS, BWS) and three PC-based competing methods (PC, nPC, sPC) to capture gene-level signals, several penalized regression approaches can be used to select genes that are related to a phenotype, including elastic net (ENET) and least absolute shrinkage and selection operator (Lasso). However, ENET and Lasso ignore genetic network structures that are expected to perform poorer according to the feature selection. To compare the network-based regression (Net) with ENET and Lasso, we use the following procedure: 1) Calculate gene-level signals using six methods. 2) Apply three different regressions, Net, Lasso, and ENET, to each of six gene-level signals. 3) Calculate selection probability based on the half-sample approach for each of the 18 combinations, which contains six gene-level signals from 1) and three regressions from 2), such as OWS+Net, OWS+Lasso, OWS+ENET, etc. 4) After we obtain the selection probabilities of each combination, we select top 100 genes and then calculate the true positive rates (TPRs).

The penalty of ENET is defined as $P_{ENET}(\boldsymbol{\beta}) = \lambda\alpha\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\boldsymbol{\beta}$ and the penalty of Net is defined as $P_{Net}(\boldsymbol{\beta}) = \lambda\alpha\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\mathbf{S}^T\mathbf{L}\mathbf{S}\boldsymbol{\beta}$. The only difference between these two penalties is the second term, where $\mathbf{S}^T\mathbf{L}\mathbf{S}$ represents the network structure among all genes. If there is no network structure among genes, all elements of the adjacency matrix $\mathbf{A}$ equal 0 since there is no pair of genes that are connected. The symmetric normalized Laplacian matrix $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = \mathbf{I}$, so the penalty of Net is equal to the penalty of ENET, that is, $P_{Net}(\boldsymbol{\beta}) = \lambda\alpha\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\mathbf{S}^T\mathbf{L}\mathbf{S}\boldsymbol{\beta} = \lambda\alpha\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda(1-\alpha)\boldsymbol{\beta}^T\mathbf{I}\boldsymbol{\beta} = P_{ENET}(\boldsymbol{\beta})$.
Therefore, ENET is a special case of Net without considering the network structure.

Figures E.9 – E.12 show the TPR comparison results for DNA methylation and DNA sequence data with balanced (**Figures E.9 and E.10**) and unbalanced (**Figures E.11 and E.12**) disease status under different simulation settings. We can observe that Net is better than Lasso and ENET in all simulation scenarios, and ENET is better than Lasso in most scenarios. That means, if there is a network structure of the genes, the network-based regression performs better regarding the selection. Meanwhile, we observe that the three weighted combinations (OWS, LD-PRS, BWS) along with Net always perform better than those along with Lasso and ENET in all simulation settings. However, the three PC-based competing methods along with Net may not increase TPR compared with Lasso and ENET.

Therefore, the methods used to capture gene-level signals are very important for feature selection. We can conclude that the performance of feature selection will be boosted if we can capture more information on the gene-level signals.

Computational efficiency is very important for analyzing high-dimensional genomic data. We compare the computational time between our proposed methods and the competing methods. To include the stability selection in the total time, we choose 600 pairs of tuning parameters to evaluate ENET and Net; 500 tuning parameters to evaluate Lasso. For each of the three regressions (Net, ENET, Lasso), there are six methods to capture the gene-level signals, eight combinations of parameters (δ, ω, σ^2) in the simulation, and two scenarios of the network structure. Based on $6 \times 8 \times 2 = 96$ replicates in the simulation, we estimate the computational time (s) of the three regressions based on 1,000 genes and 1,000 sample size. All analyses are performed on a macOS (2.7 GHz Quad-Core Intel Core i7, 16 GB memory). The computing times are shown in **Figure E.13**. Based on the figure, although network-based regressions need more computational time than ENET and Lasso, the average time of Net is only 236.75s by using the half-sample approach 100 times. In the application of the DNA methylation data, we consider 10,737 genes and 689 individuals. The average computational time for network-based regression based on 600 pairs of tuning parameters is 6 hours by using the half-sample approach 500 times. In the application to the DNA sequence data, we consider 10,907 genes and 10,000 individuals. The average computational time for the network-based regression based on 600 pairs of tuning parameters is 11 hours by using the half-sample approach 500 times.

Text E.3. Evaluation of Model Fitting.

Model validation is a most important step in the model building process, which is carried out after model training where the trained model is evaluated with a separate testing data set. To evaluate the model fitting regarding our proposed methods along with the selection probability, we use the following steps. First, we calculate the selection probabilities of all genes by using the methods with three weighted combinations (OWS, LD-PRS, BWS) as well as using the three competing methods (PC, nPC, sPC). Then, we choose the top K genes with the largest selection probabilities from each of the six methods. We use the accuracy rate (ACC) as a measurement to evaluate models, where ACC is defined as the sum of true positives and true negatives divided by the total number of genes[49]. Based on different numbers of top selected genes, we perform a 10-fold cross-validation to calculate the average ACC and the standard deviation of ACC.

There are 45 genes in scenario 1 and 48 genes in scenario 2 that are related to the phenotype. To evaluate the model fitting, we select top 40 and 60 genes ($K = 40$ and 60) by each of the six methods according to the selection probabilities. **Figure E.15** shows ACCs with the standard deviations for both DNA sequence and DNA methylation data under different simulation settings for the phenotype with a balanced case-control ratio. As expected, the two proposed supervised methods, LD-PRS and OWS, have higher ACC compared with the unsupervised method, BWS. Also, sPC is the supervised competing method, which also has higher ACC compared with the other two unsupervised PC-based methods, PC and nPC. Notably, LD-PRS and OWS always outperform sPC even though all of the three methods are supervised. Among the unsupervised models, BWS has higher

ACC than the two PC-based methods in the DNA sequence data analysis, while these three methods perform equivalently in the DNA methylation data analysis.

Text E.4. Comparison of the Methods with Partially Corrected Network Structure.

One of the most important issues for applying the network-based regression is how to select a biological network. In the study, we consider the functional relationships among genes in the genetic network, which can be obtained from the existing annotations. For example, in the real data application, we construct an association network using the pathways from seven genetic network databases, where the genes are associated with each other if they are within a metabolic pathway or a biological process. However, the constructed network may contain some incorrect or missing connections between genes. Therefore, we perform the simulation studies to evaluate if the network-based regression with partially corrected network structure still outperform the regressions without considering the network structure. In the simulation studies, we consider a total of 1,000 genes with 10 subnetwork modules shown in **Figure E.2**, where the existing genetic network contains 990 undirected edges. To mimic a partially correct network structure, we randomly remove 250 out of 990 edges and then randomly add 250 new edges to the genetic network. That means the genetic network has around 25% incorrect relationships. **Figure E.14** shows the TPR results. We can see that the TPRs of the methods with the partially correct genetic network are lower than those with the correct network as expected. Even though the genetic network is partially incorrect, the performance of the network-based regressions still performs better than the methods without considering the genetic network structure (ENET and Lasso) in most simulation settings.

Text E.5. Real Data Applications for DNA Methylation and DNA Sequence

Biological network: In order to utilize biological network information, we employ the same pathway information as in Kim et al.[45] , where there are seven genetic network databases from Biocarta, HumnaCyc, KEGG, NCI, Panther, Reactome, and SPIKE in R package 'graphite'. There are 11,381 linked genes in the package, of which 672,571 edges among those genes are in the biological network. To matach SNPs and CpG sites to the linked genes, we consider all genes according to the USCS (GRCh37/hg19) genome sequence annotation list which can be downloaded from the UCSC website (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/).

DNA methylation data: The DNA methylation data was measured using the Illumina HumanMethylation450 BeadChip from 354 RA patients (cases) and 335 normal controls[50,51]. The dataset can be found in the NCBI Gene Expression Omnibus (GEO) with identifier GSE42861, where the methylation β values of CpG sites are provided. Then, we convert β values to M values using logit (base 2) function for further analysis. We find that only 10,737 linked genes matched with genes in the above biological network that contain at least one CpG site. After pruning CpG sites in each gene, we capture gene-level signals using OWS, LD-PRS, BWS, and nPC.

DNA sequence data: The UK Biobank is a population-based cohort study with a wide variety of genetic and phenotypic information[52]. It includes ~500K individuals from the United Kingdom who are currently aged between 40 and 69 when recruited in 2006-2010[53]. We follow the same preprocess procedure in Liang et al.[54] to exclude individuals who self-

report themselves not from a white British ancestry, who are marked as outliers for heterozygosity or missing rates, who have been identified to have ten or more third-degree relatives, and who are recommended for removal by the UK Biobank. Meanwhile, the quality control (QC) for DNA sequence data is also performed on both SNPs and samples using PLINK 1.9[55] (https://www.cog-genomics.org/plink/1.9/). We filter out SNPs with missing rates larger than 5% and Hardy-Weinberg equilibrium exact test p-values less than 10^{-6}. We also exclude individuals with missing rates larger than 5% and without sex. After QC of DNA sequence data and preprocess of phenotype data, there are 583,386 SNPs and 322,607 individuals remaining. In our analysis, we use 4,541 individuals with RA disease and randomly select 5,459 individuals without RA disease. We define a gene to include all of the SNPs from 20 kb upstream to 20 kb downstream of the gene. We find that only 10,907 linked genes matched with genes in the above biological network that contain at least one SNP. Then, we capture gene-level signals using OWS, LD-PRS, BWS, and nPC, respectively.

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways: In this study, we map those genes to the KEGG pathways using a functional annotation tool named Database for Annotation, Visualization, and Integrated Discovery Bioinformatics Resource[56,57] (DAVID: https://david.ncifcrf.gov/) for pathway enrichment analysis.

E.2 Supplementary Tables

Table E.1. AUCs of the six methods for DNA sequence data analyses in all simulation settings.

Method	$\delta = 2$ $\omega = 4$ $\sigma^2 = 2$	$\delta = 2$ $\omega = 4$ $\sigma^2 = 3$	$\delta = 2$ $\omega = 6$ $\sigma^2 = 2$	$\delta = 2$ $\omega = 6$ $\sigma^2 = 3$	$\delta = 3$ $\omega = 4$ $\sigma^2 = 2$	$\delta = 3$ $\omega = 4$ $\sigma^2 = 3$	$\delta = 3$ $\omega = 6$ $\sigma^2 = 2$	$\delta = 3$ $\omega = 6$ $\sigma^2 = 3$
Scenario 1: DNA sequence (Balance)								
OWS	0.77	0.56	0.98	**0.94**	0.86	0.76	**0.99**	**0.99**
LD-PRS	0.84	0.60	**0.99**	0.93	0.89	0.84	**0.99**	**0.99**
BWS	**0.88**	**0.74**	0.96	0.89	**0.97**	**0.92**	**0.99**	**0.99**
PC	0.50	0.56	0.67	0.62	0.54	0.50	0.84	0.66
nPC	0.53	0.57	0.69	0.64	0.57	0.54	0.84	0.69
sPC	0.48	0.58	0.67	0.64	0.51	0.52	0.81	0.65
Scenario 2: DNA sequence (Balance)								
OWS	0.79	0.67	**0.96**	**0.97**	0.86	0.86	**0.99**	**0.99**
LD-PRS	0.84	0.68	**0.96**	**0.96**	0.89	0.89	**0.99**	**0.99**
BWS	**0.86**	**0.82**	0.92	0.90	**0.97**	**0.91**	0.98	**0.99**
PC	0.52	0.49	0.71	0.65	0.55	0.56	0.81	0.78
nPC	0.53	0.52	0.73	0.61	0.54	0.54	0.78	0.77
sPC	0.53	0.59	0.66	0.54	0.54	0.55	0.80	0.74
Scenario 1: DNA sequence (Unbalance)								
OWS	0.66	0.62	**0.93**	**0.85**	0.76	0.71	**0.95**	**0.89**
LD-PRS	0.67	0.64	**0.93**	0.82	0.76	0.71	0.94	**0.89**
BWS	**0.83**	**0.75**	0.89	0.74	**0.86**	**0.79**	**0.95**	0.87
PC	0.55	0.53	0.73	0.69	0.55	0.54	0.79	0.60
nPC	0.54	0.51	0.74	0.66	0.57	0.55	0.79	0.61
sPC	0.53	0.54	0.76	0.69	0.54	0.53	0.81	0.63
Scenario 2: DNA sequence (Unbalance)								
OWS	0.68	0.60	**0.89**	**0.83**	0.80	0.65	**0.92**	**0.97**
LD-PRS	0.70	0.60	0.88	0.81	0.75	0.66	0.91	0.96
BWS	**0.78**	**0.74**	0.85	0.73	**0.88**	**0.78**	0.87	0.85
PC	0.50	0.53	0.68	0.62	0.53	0.58	0.76	0.67
nPC	0.51	0.56	0.67	0.58	0.54	0.54	0.76	0.66
sPC	0.49	0.54	0.68	0.60	0.49	0.58	0.73	0.69

Note: the bold-faced values denote the maximum AUC across all six methods in the same simulation settings.

Table E.2. AUCs of the six methods in DNA methylation data analyses in all simulation settings.

Method	$\delta = 2$ $\omega = 4$ $\sigma^2 = 6$	$\delta = 2$ $\omega = 4$ $\sigma^2 = 7$	$\delta = 2$ $\omega = 6$ $\sigma^2 = 6$	$\delta = 2$ $\omega = 6$ $\sigma^2 = 7$	$\delta = 2.5$ $\omega = 4$ $\sigma^2 = 6$	$\delta = 2.5$ $\omega = 4$ $\sigma^2 = 7$	$\delta = 2.5$ $\omega = 6$ $\sigma^2 = 6$	$\delta = 2.5$ $\omega = 6$ $\sigma^2 = 7$
Scenario 1: DNA methylation (Balance)								
OWS	0.91	0.86	0.96	0.93	0.97	**0.95**	0.97	0.96
LD-PRS	0.90	0.86	0.96	0.94	0.97	0.93	0.97	0.96
BWS	**0.93**	**0.89**	**0.97**	**0.95**	**0.99**	0.94	**0.99**	**0.99**
PC	0.76	0.75	0.95	0.85	0.88	0.74	0.94	0.90
nPC	0.78	0.76	0.95	0.84	0.86	0.75	0.94	0.91
sPC	0.80	0.76	0.90	0.80	0.89	0.70	0.93	0.92
Scenario 2: DNA methylation (Balance)								
OWS	0.91	0.90	0.95	0.94	**0.97**	0.92	0.98	0.96
LD-PRS	0.90	0.91	0.96	0.93	0.96	0.90	0.98	0.97
BWS	**0.94**	**0.92**	**0.98**	**0.96**	**0.97**	**0.92**	**1.00**	**0.98**
PC	0.85	0.79	0.95	0.87	0.85	0.77	0.93	0.89
nPC	0.85	0.79	0.96	0.87	0.84	0.79	0.92	0.91
sPC	0.79	0.74	0.95	0.85	0.80	0.75	0.93	0.90
Scenario 1: DNA methylation (Unbalance)								
OWS	**0.81**	**0.81**	0.90	0.90	**0.90**	**0.86**	0.94	0.88
LD-PRS	0.80	0.80	0.89	**0.91**	0.89	0.85	**0.95**	0.87
BWS	**0.81**	**0.81**	**0.94**	0.88	0.80	0.80	0.93	**0.95**
PC	0.66	0.60	0.82	0.77	0.69	0.70	0.81	0.74
nPC	0.64	0.59	0.81	0.78	0.69	0.72	0.81	0.75
sPC	0.66	0.62	0.80	0.63	0.67	0.59	0.71	0.68
Scenario 2: DNA methylation (Unbalance)								
OWS	**0.87**	**0.77**	0.89	0.84	**0.89**	0.87	**0.96**	0.87
LD-PRS	**0.87**	0.76	0.91	0.83	**0.89**	**0.88**	**0.96**	0.88
BWS	0.80	0.69	**0.93**	**0.91**	0.88	0.86	**0.96**	**0.94**
PC	0.79	0.58	0.84	0.68	0.70	0.69	0.83	0.80
nPC	0.75	0.57	0.82	0.70	0.70	0.70	0.85	0.79
sPC	0.61	0.62	0.73	0.64	0.69	0.56	0.74	0.71

Note: the bold-faced values denote the maximum AUC across all six methods in the same simulation settings.

E.3 Supplementary Figures

Figure E.1. The graphical abstract of the methods in this study. To capture gene-level signals from multiple CpG sites or SNPs in the m^{th} gene ($m=1,\cdots,M$), we first employ three weighted combinations, OWS, BWS, and LD-PRS (left). Then we use half-sample approach B times on the phenotype ($\boldsymbol{y}$) and gene-level signals ($\boldsymbol{z}_1,\cdots,\boldsymbol{z}_M$) (center). For each time of the half-sample approach, we apply the network-based reression for a grid set of tuning parameters. Finally, we calculate the selection probability of each gene and select genes with the highest selection probabilities (right).

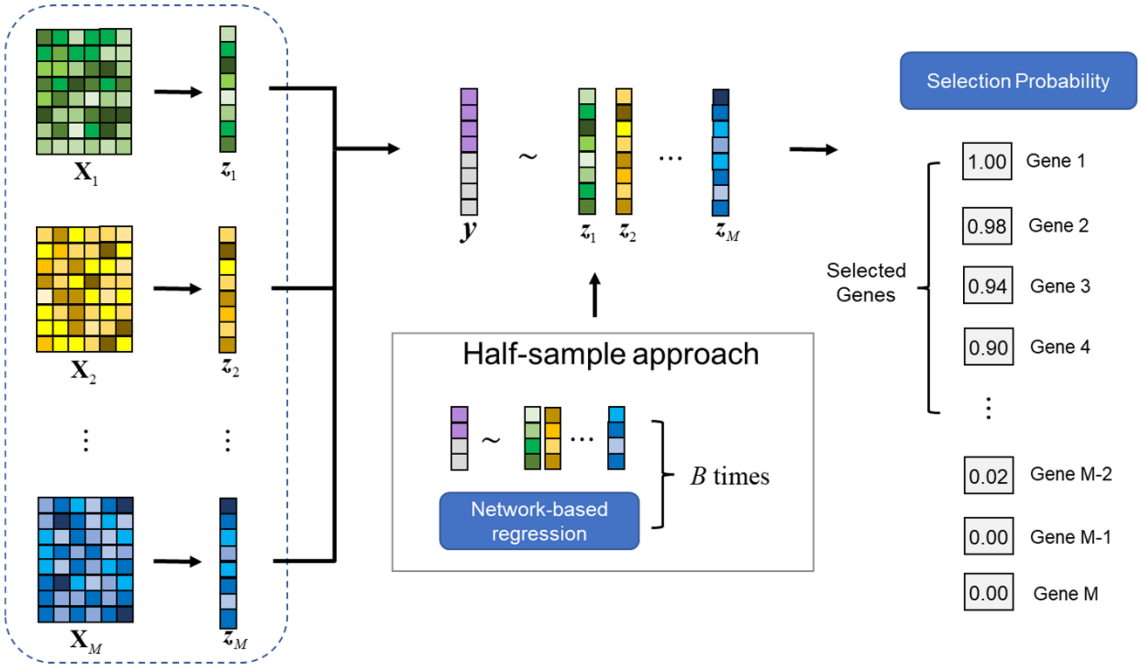


Figure E.2. The network module used in simulation studies. There are a total of 100 genes which contain a centered gene.

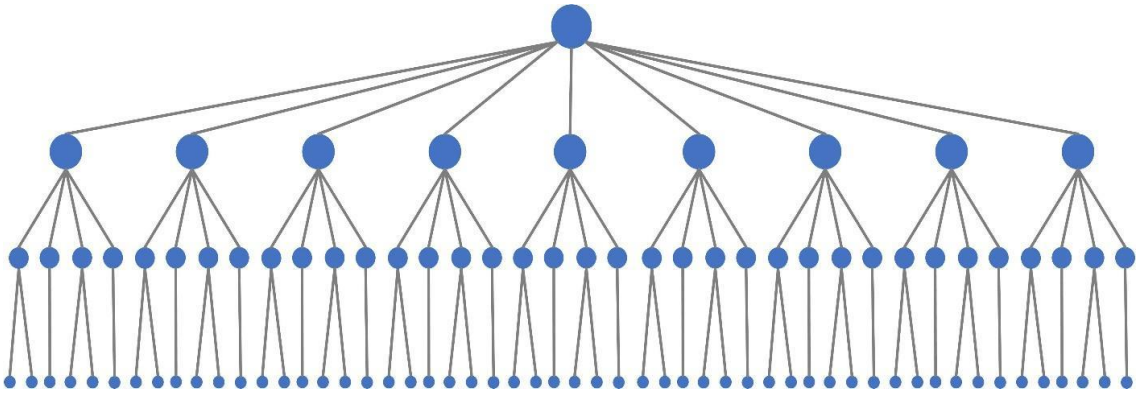


Figure E.3. The true positive rates of the methods based on different gene-level signals for balance case-control studies with DNA sequence data in scenario 2, where there are five rare variants and five common variants in each gene. According to the different number of selected top-genes, three parameters are used to vary the genetic effect: the strength of association signals δ, the number of SNPs in each gene related to gene-level signals ω, and the noise level of association signals σ^2. The selection probabilities are calculated using half-sample method 100 times.

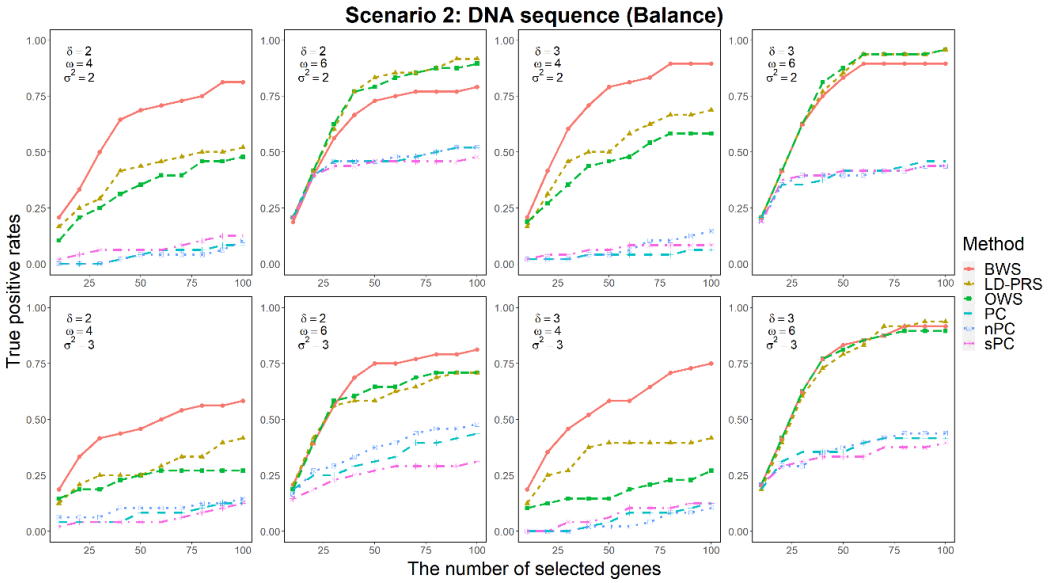


Figure E.4. The true positive rates of the methods based on different gene-level signals for balance case-control studies with DNA methylation data in scenario 2. According to the different number of selected top-genes, three parameters are used to vary the genetic effect: the strength of association signals δ, the number of CpG sites in each gene related to gene-level signals ω, and the noise level of association signals σ^2. The selection probabilities are calculated using half-sample method 100 times.

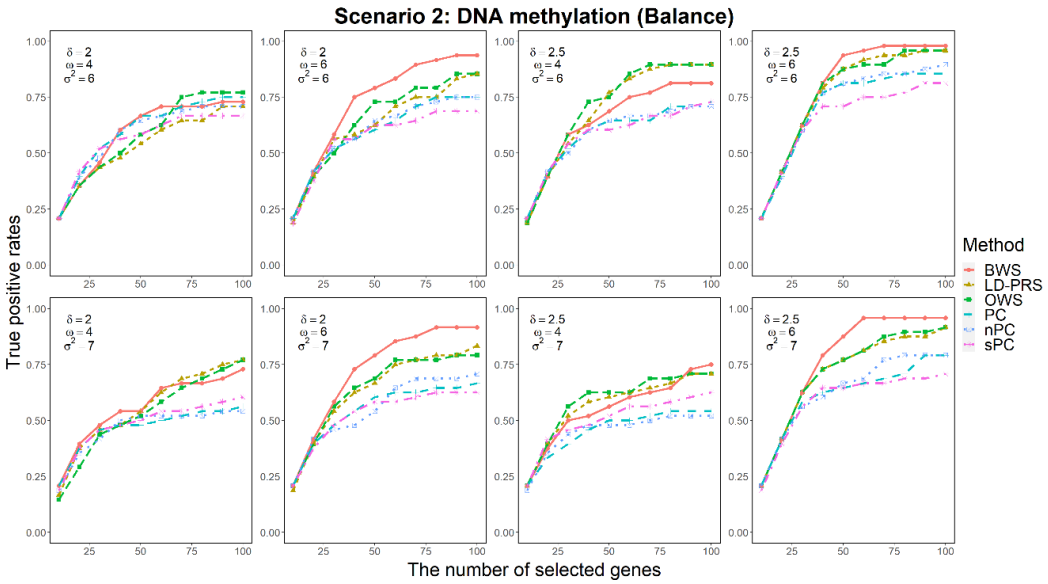


Figure E.5. The true positive rates of the methods based on different gene-level signals for unbalance case-control studies (case:control=100:900) with DNA sequence data in scenario 1, where there are five rare variants and five common variants in each gene. According to the different number of selected top-genes, three parameters are used to vary the genetic effect: the strength of association signals δ, the number of SNPs in each gene related to gene-level signals ω, and the noise level of association signals σ^2. The selection probabilities are calculated using half-sample method 100 times.

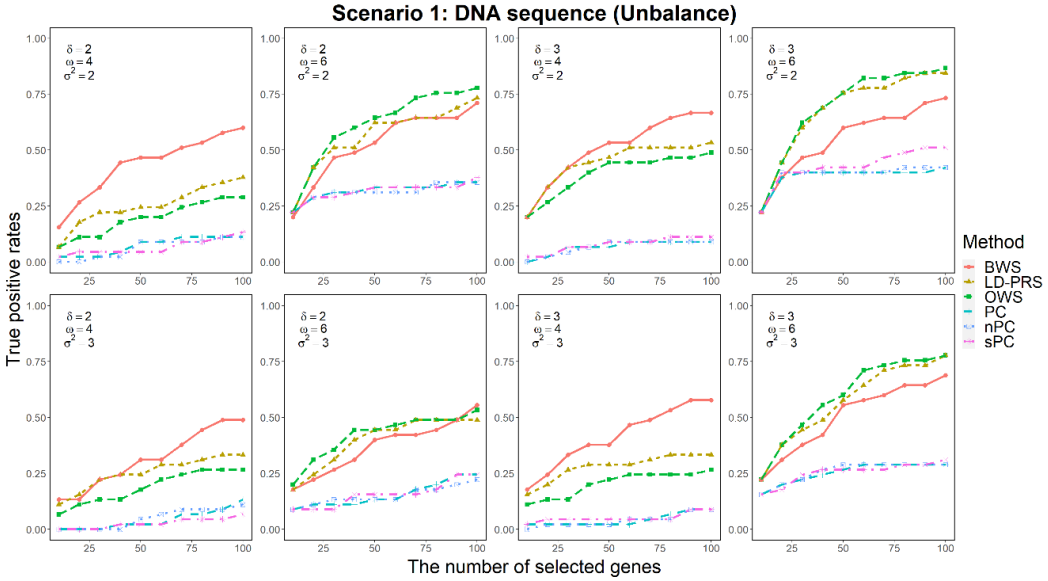


Figure E.6. The true positive rates of the methods based on different gene-level signals for unbalance case-control studies (case:control=100:900) with DNA methylation data in scenario 1. According to the different number of selected top-genes, three parameters are used to vary the genetic effect: the strength of association signals δ, the number of CpG sites in each gene related to gene-level signals ω, and the noise level of association signals σ^2. Selection probabilities are calculated using half-sample method 100 times.

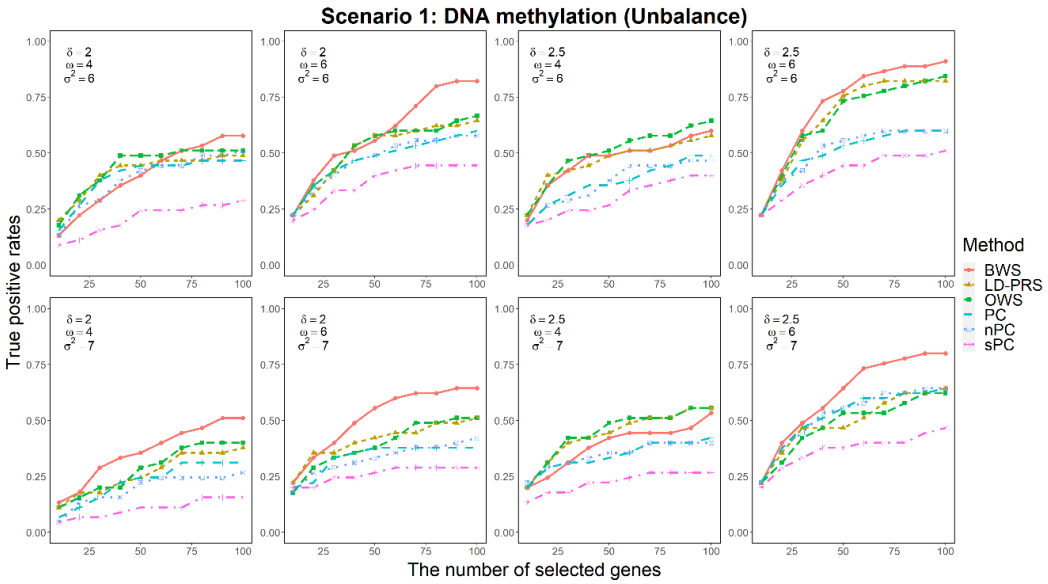


Figure E.7. The true positive rates of the methods based on different gene-level signals for unbalance case-control studies (case:control=100:900) with DNA sequence data in scenario 2, where there are five rare variants and five common variants in each gene. According to the different number of selected top-genes, three parameters are used to vary the genetic effect: the strength of association signals δ, the number of SNPs in each gene related to gene-level signals ω, and the noise level of association signals σ^2. The selection probabilities are calculated using half-sample method 100 times.

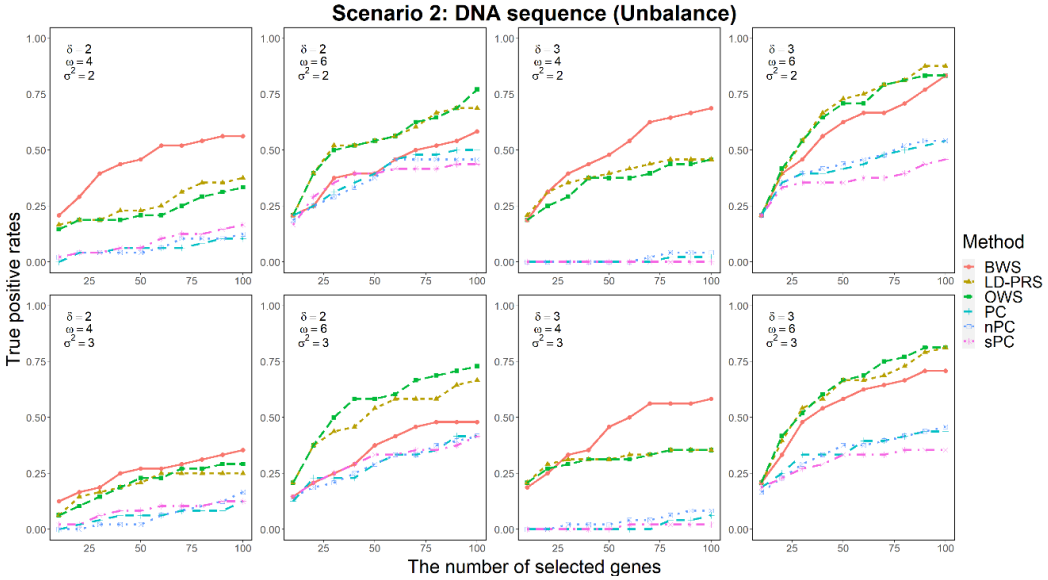


Figure E.8. The true positive rates of the methods based on different gene-level signals for unbalance case-control studies (case:control=100:900) with DNA methylation data in scenario 2. According to the different number of selected top-genes, three parameters are used to vary the genetic effect: the strength of association signals δ, the number of CpG sites in each gene related to gene-level signals ω, and the noise level of association signals σ^2. Selection probabilities are calculated using half-sample method 100 times.

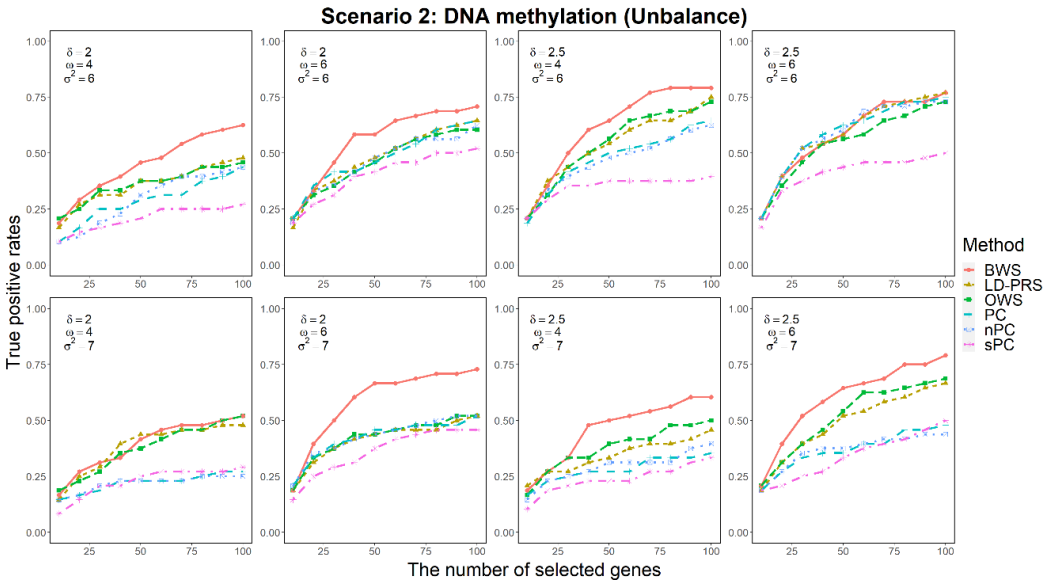


Figure E.9. The true positive rates of the methods by selected top 100 genes according to the selection probabilities based on different gene-level signals for balance case-control studies with DNA sequence data in scenarios 1 and 2. The selection probabilities are calculated using half-sample method 100 times.

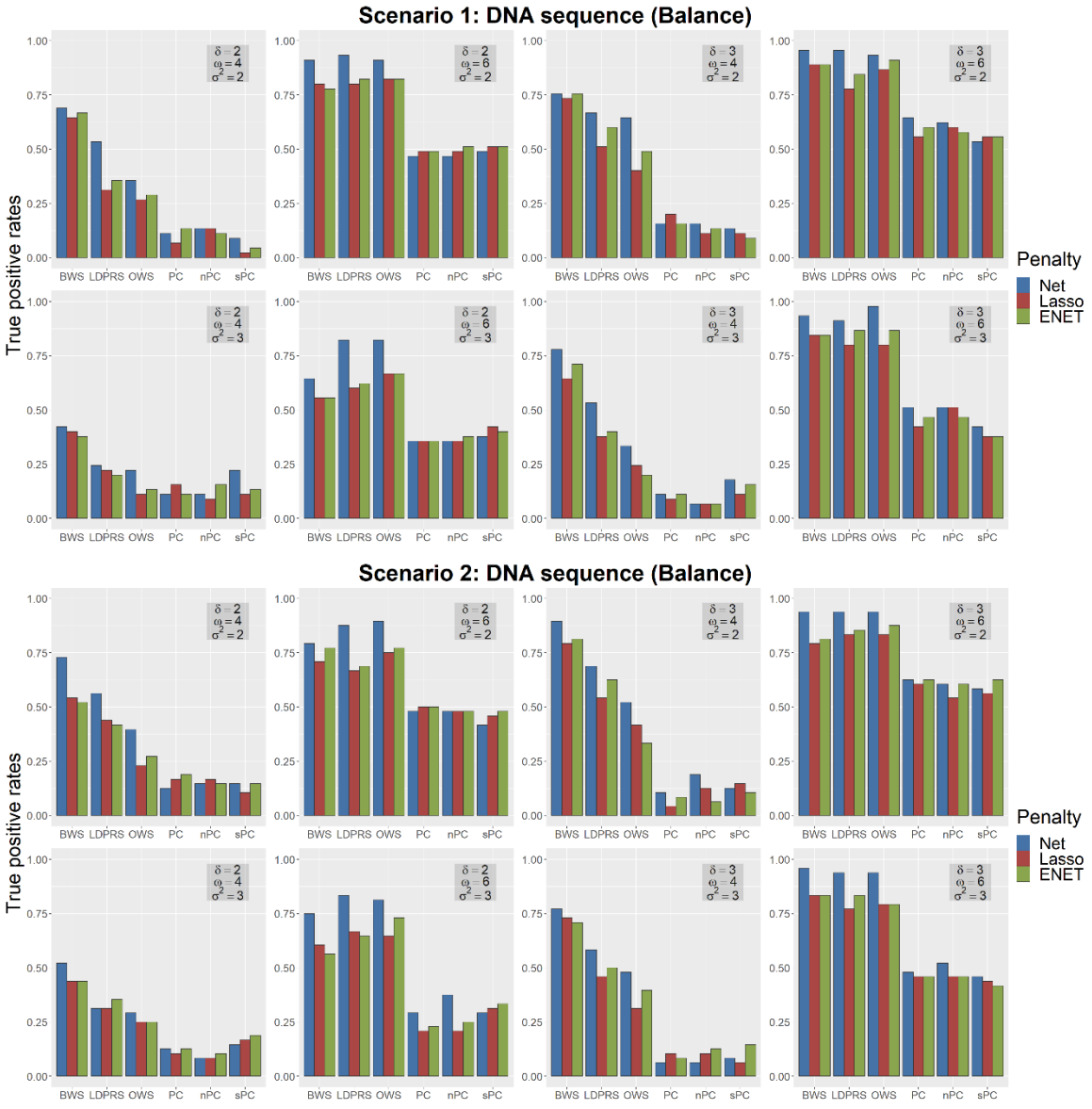


Figure E.10. The true positive rates of the methods by selected top 100 genes according to the selection probabilities based on different gene-level signals for balance case-control studies with DNA methylation data in scenarios 1 and 2. The selection probabilities are calculated using half-sample method 100 times.

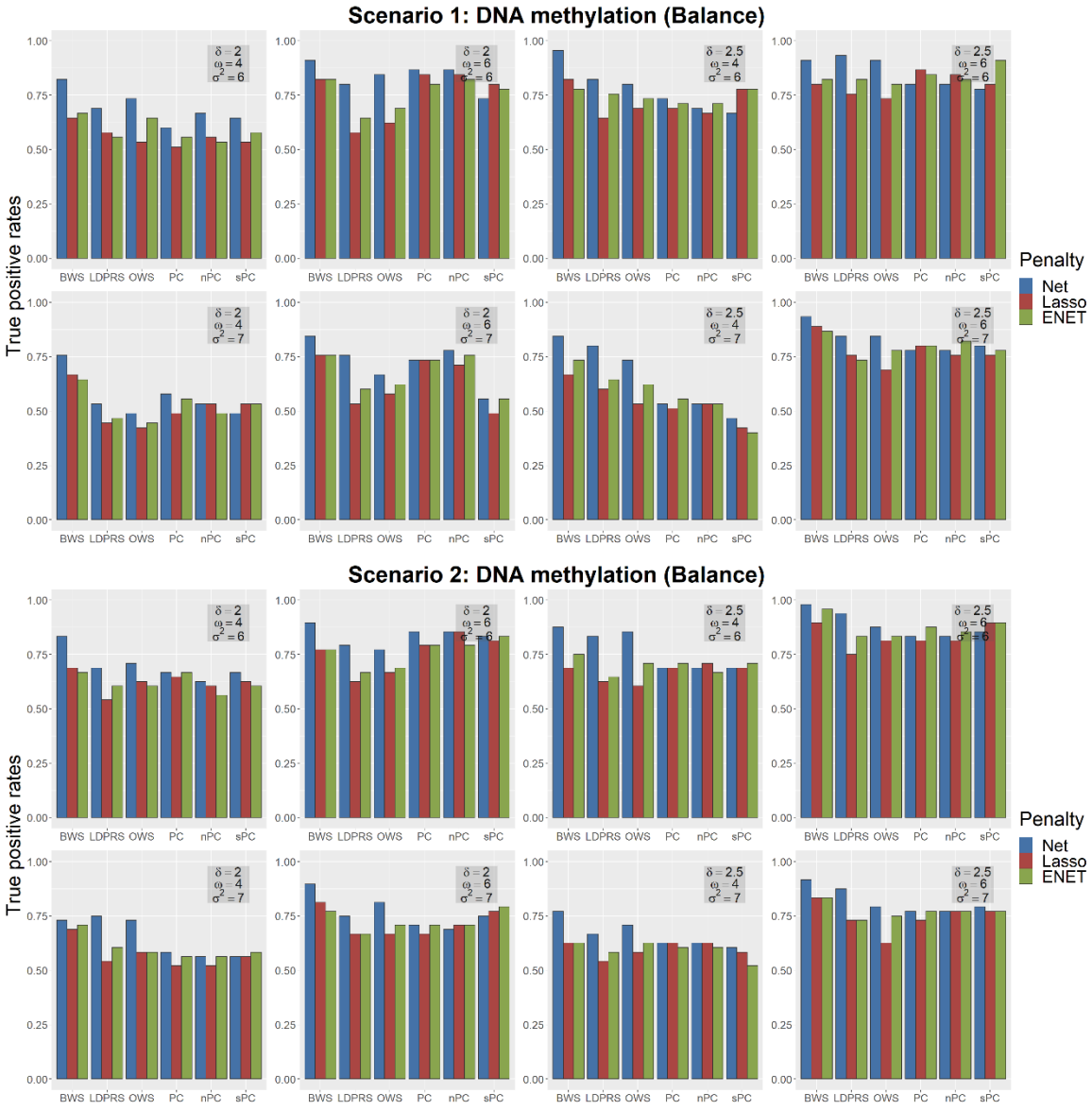


Figure E.11. The true positive rates of the methods by selected top 100 genes according to the selection probabilities based on different gene-level signals for unbalance case-control studies (case:control=100:900) with DNA sequence data in scenarios 1 and 2. The selection probabilities are calculated using half-sample method 100 times.

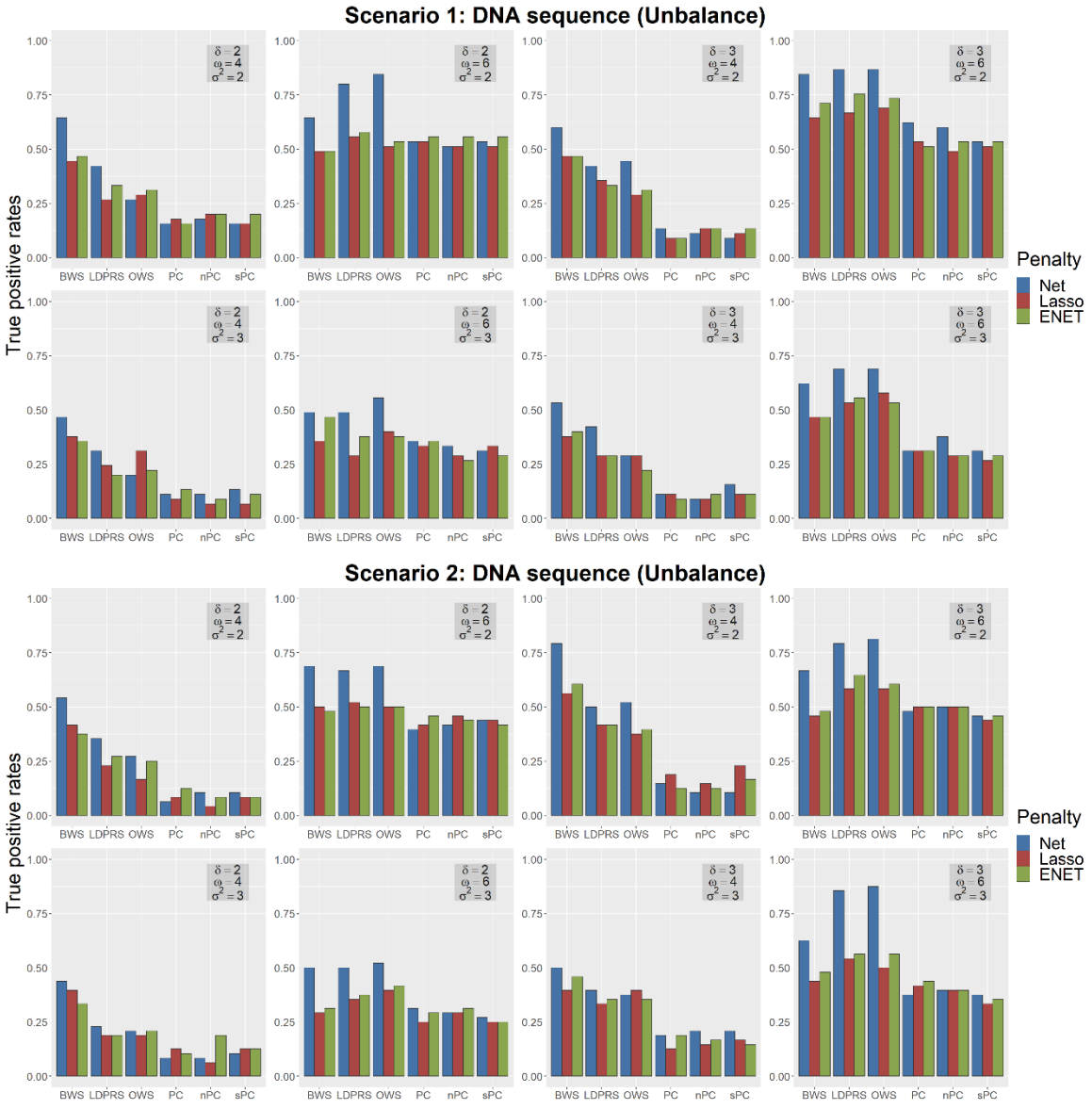


Figure E.12. The true positive rates of the methods by selected top 100 genes according to the selection probabilities based on different gene-level signals for unbalance case-control studies (case:control=100:900) with DNA methylation data in scenarios 1 and 2. The selection probabilities are calculated using half-sample method 100 times.

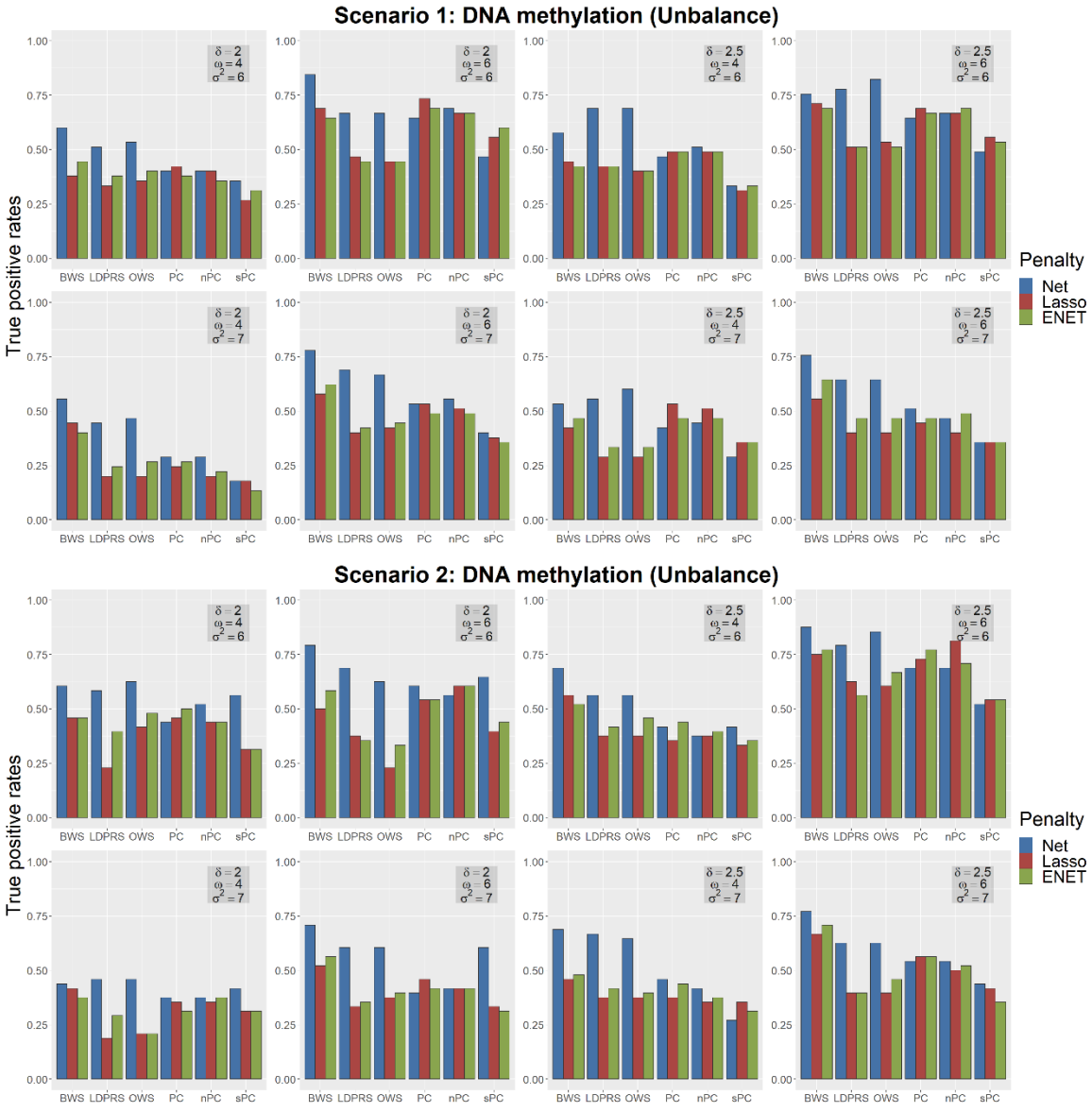


Figure E.13. The comparison of computational time required by the methods with Net regression and the methods with two methods without considering the network structure (ENET and Lasso). We choose 600 pairs of tuning parameters to evaluate ENET and Net; 500 tuning parameters to evaluate Lasso; 100 times of the half-sample approach.

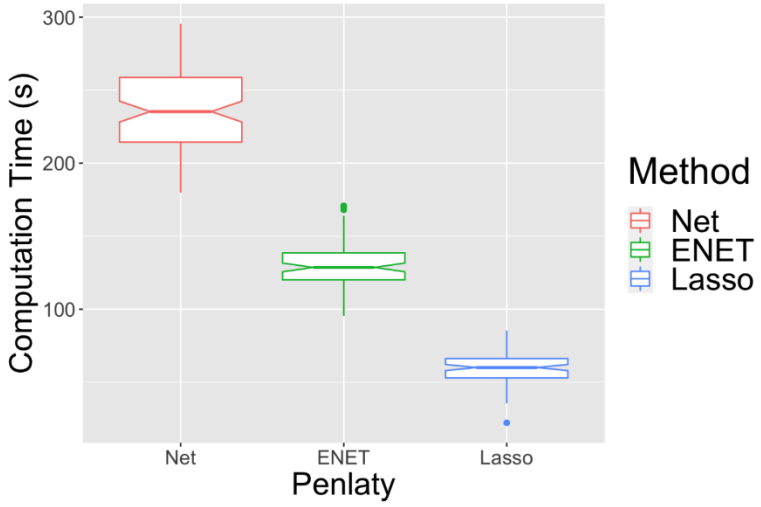


Figure E.14. The accuracy (ACC) with standard deviation for both DNA sequence and DNA methylation data under different simulation settings for the phenotype with a balanced case-control ratio.

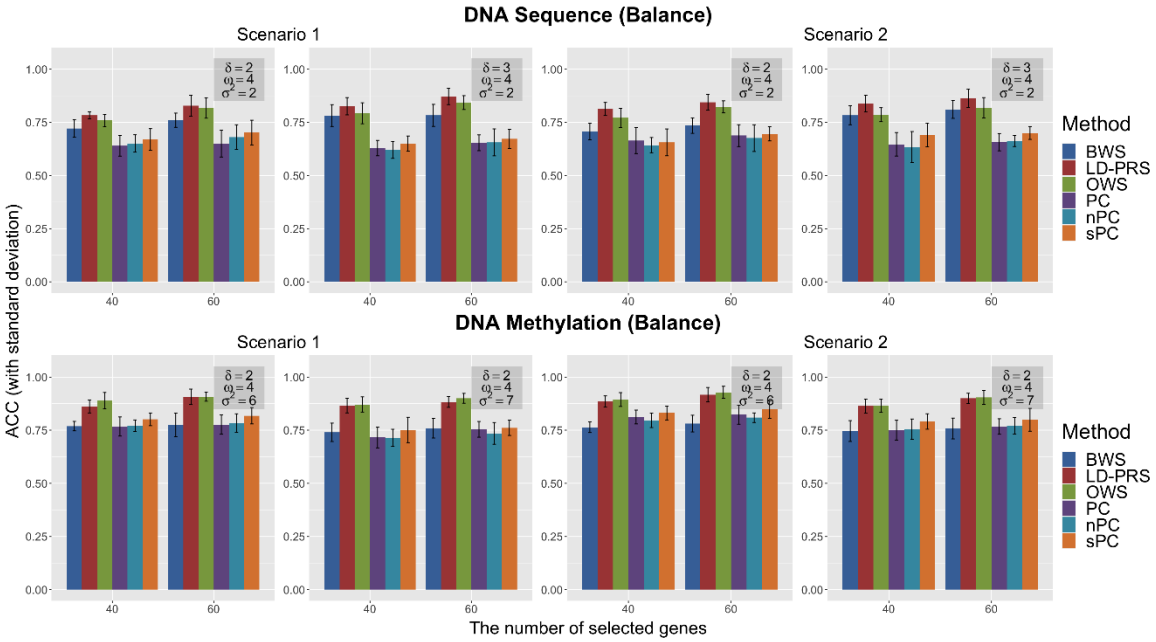


Figure E.15. The true positive rates of the methods by selected top 100 genes according to the selection probabilities based on different gene-level signals for balance case-control studies with DNA sequence data in scenario 2. The selection probabilities are calculated using half-sample method 100 times.

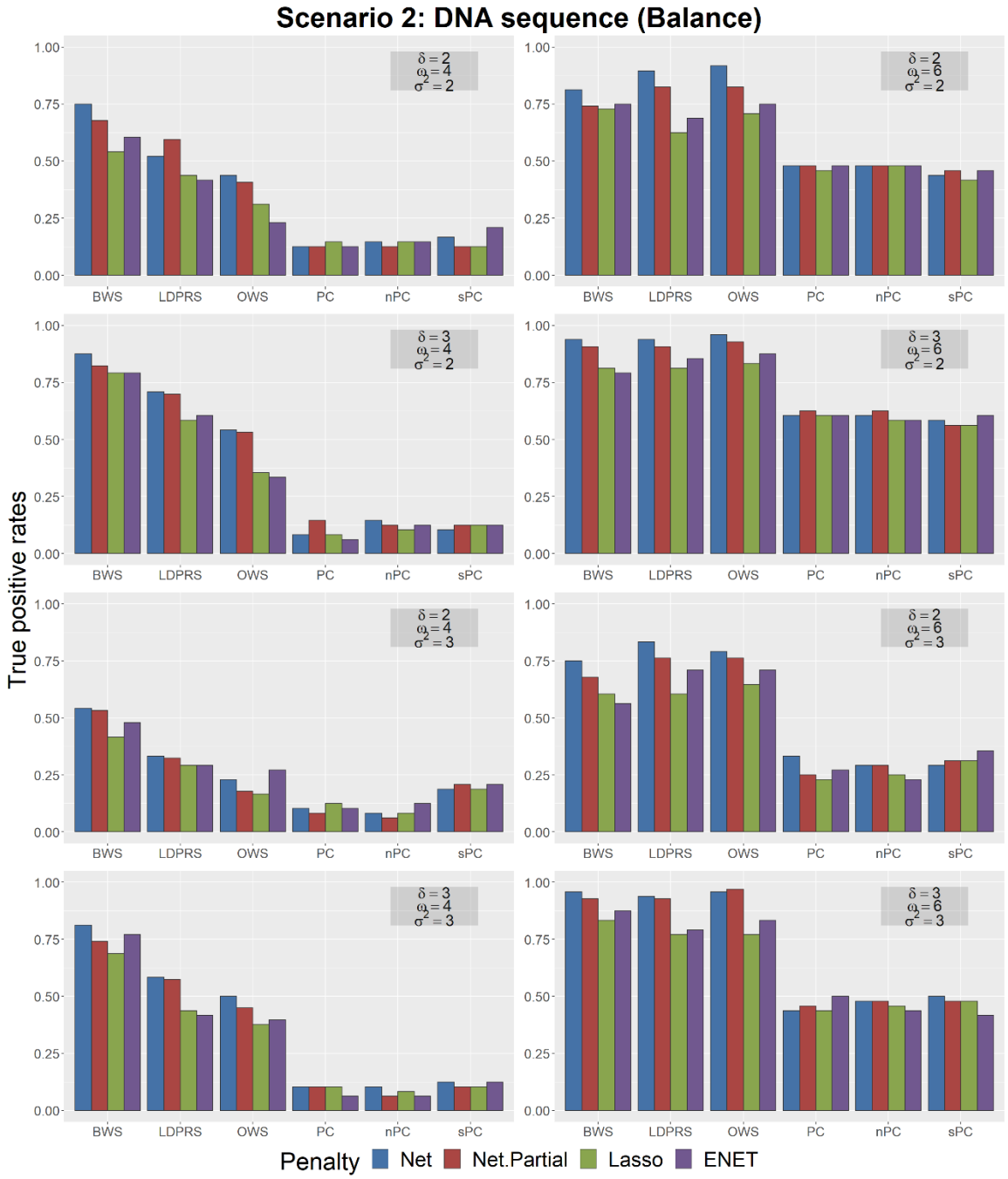


Figure E.16. Venn diagram of the number of top 100 genes identified by BWS, LD-PRS, OWS, and nPC for DNA methylation data.

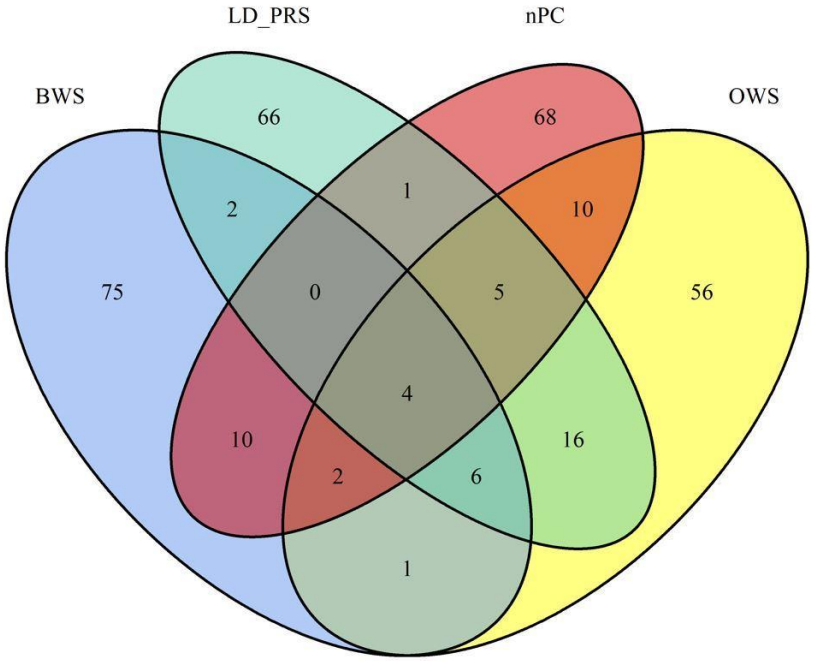
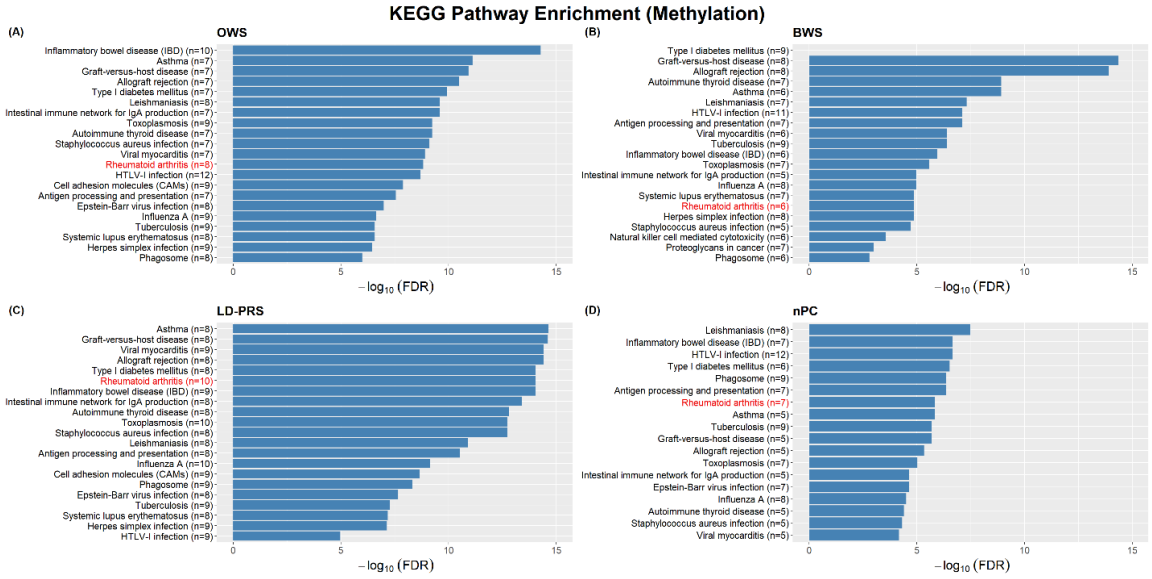


Figure E.17. The KEGG pathway enrichment analysis results of BWS, LD-PRS, OWS, and nPC for DNA methylation data.



F Reference List for Supplementary Materials

1 Sha, Q., Wang, Z., Zhang, X. & Zhang, S. A clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. *Bioinformatics* **35**, 1373-1379 (2019).

2 Wang, M., Zhang, S. & Sha, Q. A computationally efficient clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. *PloS one* **17**, e0260911 (2022).

3 Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association* **115**, 393-402 (2020).

4 Liu, Y. *et al.* ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics* **104**, 410-421 (2019).

5 Liang, X., Cao, X., Sha, Q. & Zhang, S. HCLC-FC: A novel statistical method for phenome-wide association studies. *Plos one* **17**, e0276646 (2022).

6 Liang, X., Sha, Q., Rho, Y. & Zhang, S. A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. *Genetic epidemiology* **42**, 344-353 (2018).

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

8 O'Brien, P. C. Procedures for comparing samples with multiple endpoints. *Biometrics*, 1079-1087 (1984).

9 Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *The American Journal of Human Genetics* **101**, 37-49 (2017).

10 Tachmazidou, I. *et al.* Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nature genetics* **51**, 230-236 (2019).

11 Kim, S. K., Nguyen, C., Jones, K. B. & Tashjian, R. Z. A Genome Wide Association Study For Shoulder Impingement and Rotator Cuff Disease. *Journal of Shoulder and Elbow Surgery* (2021).

12 Johnston, K. J. *et al.* Genome-wide association study of multisite chronic pain in UK Biobank. *PLoS genetics* **15**, e1008164 (2019).

13 Aterido, A. *et al.* Genetic variation at the glycosaminoglycan metabolism pathway contributes to the risk of psoriatic arthritis but not psoriasis. *Annals of the Rheumatic diseases* **78**, 355-364 (2019).

14 Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature genetics* **47**, 1457-1464 (2015).

15 Kim, S. K. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. *PloS one* **13**, e0200785 (2018).

16 Hou, S. *et al.* Identification of a susceptibility locus in STAT4 for Behçet's disease in Han Chinese in a genome‐wide association study. *Arthritis & Rheumatism* **64**, 4104-4113 (2012).

17 Renauer, P. A. *et al.* Identification of susceptibility loci in IL6, RPS9/LILRB3, and an intergenic locus on chromosome 21q22 in Takayasu arteritis in a genome‐wide association study. *Arthritis & rheumatology* **67**, 1361-1368 (2015).
18 Allanore, Y. *et al.* Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis. *PLoS Genet* **7**, e1002091 (2011).
19 Chung, S. A. *et al.* Lupus nephritis susceptibility loci in women with systemic lupus erythematosus. *Journal of the American Society of Nephrology* **25**, 2859-2870 (2014).
20 Rothwell, S. *et al.* Dense genotyping of immune-related loci in idiopathic inflammatory myopathies confirms HLA alleles as the strongest genetic risk factor and suggests different genetic background for major clinical subgroups. *Annals of the rheumatic diseases* **75**, 1558-1566 (2016).
21 Miller, F. W. *et al.* Genome-wide association study identifies HLA 8.1 ancestral haplotype alleles as major genetic risk factors for myositis phenotypes. *Genes & Immunity* **16**, 470-480 (2015).
22 Cordero, A. I. H. *et al.* Genome-wide associations reveal human-mouse genetic convergence and modifiers of myogenesis, CPNE1 and STC2. *The American Journal of Human Genetics* **105**, 1222-1236 (2019).
23 Johnson, T. A. *et al.* Association of an IGHV3-66 gene variant with Kawasaki disease. *Journal of human genetics* **66**, 475-489 (2021).
24 Terao, C. *et al.* The human AIRE gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. *Human molecular genetics* **20**, 2680-2685 (2011).
25 Freitas, C. G., Aquino, A. L., Ramos, H. S., Frery, A. C. & Rosso, O. A. A detailed characterization of complex networks using Information Theory. *Scientific reports* **9**, 16689 (2019).
26 Kim, J., Bai, Y. & Pan, W. An adaptive association test for multiple phenotypes with GWAS summary statistics. *Genetic epidemiology* **39**, 651-663 (2015).
27 Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics* **50**, 229-237 (2018).
28 Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *The American Journal of Human Genetics* **96**, 21-36 (2015).
29 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47**, 1236-1241 (2015).
30 Lee, C. H., Shi, H., Pasaniuc, B., Eskin, E. & Han, B. PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics. *The American Journal of Human Genetics* **108**, 36-48 (2021).
31 Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics* **51**, 63-75 (2019).
32 Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature genetics* **51**, 237-244 (2019).

33 Grove, J. *et al.* Common risk variants identified in autism spectrum disorder. *biorxiv*, 224774 (2017).
34 Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature genetics* **48**, 624-633 (2016).
35 Meier, S. M. *et al.* Genetic variants associated with anxiety and stress-related disorders: a genome-wide association study and mouse-model study. *JAMA psychiatry* **76**, 924-932 (2019).
36 Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience* **22**, 343-352 (2019).
37 Arnold, P. D. *et al.* Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Molecular psychiatry* **23**, 1181-1181 (2018).
38 Watson, H. J. *et al.* Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nature genetics* **51**, 1207-1214 (2019).
39 Stahl, E. A. *et al.* Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics* **51**, 793-803 (2019).
40 Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics* **50**, 381-389 (2018).
41 Lee, J. J. *et al.* Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nature genetics* **50**, 1112 (2018).
42 Zhang, Y. *et al.* SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome biology* **22**, 1-30 (2021).
43 Peng, J., Wang, P., Zhou, N. & Zhu, J. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104**, 735-746 (2009).
44 Cao, X., Liang, X., Zhang, S. & Sha, Q. Gene selection by incorporating genetic networks into case-control association studies. *bioRxiv* (2022).
45 Kim, K. & Sun, H. Incorporating genetic networks into case-control association studies with high-dimensional DNA methylation data. *BMC bioinformatics* **20**, 1-15 (2019).
46 Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics* **74**, 47 (2002).
47 Parikh, N. & Boyd, S. Proximal algorithms. *Foundations and Trends in optimization* **1**, 127-239 (2014).
48 Polson, N. G., Scott, J. G. & Willard, B. T. Proximal algorithms in statistics and machine learning. *Statistical Science* **30**, 559-581 (2015).
49 Kuhn, M. & Johnson, K. *Applied predictive modeling*. Vol. 26 (Springer, 2013).
50 Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology* **31**, 142-147 (2013).

51 Kular, L. *et al.* DNA methylation as a mediator of HLA-DRB1* 15: 01 and a protective variant in multiple sclerosis. *Nature communications* **9**, 1-15 (2018).
52 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
53 Sudlow, C. *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**, e1001779 (2015).
54 Liang, X., Cao, X., Sha, Q. & Zhang, S. HCLC-FC: a novel statistical method for phenome-wide association studies. *PLoS ONE* **17(11)**, e0276646 (2022).
55 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-13015-10047-13748 (2015).
56 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1-13 (2009).
57 Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57 (2009).