Inferring a directed acyclic graph of phenotypes from GWAS summary statistics*

Rachel Zilinskas[1], Chunlin Li[2], Xiaotong Shen[2], Wei Pan[1], Tianzhong Yang[1]

February 11, 2023

Abstract

Estimating phenotype networks is a growing field in computational biology. It helps deepen the understanding of disease etiology and is useful in many applications. In this study, we present a method that constructs a phenotype network by assuming a Gaussian linear structure model embedding a directed acyclic graph (DAG). We utilize genetic variants as instrumental variables and show how our method only requires access to summary statistics from a genome-wide association study (GWAS) and a reference panel of genotype data. Besides estimation, a distinct feature of the method is its summary statistics-based likelihood ratio test on directed edges. We applied our method to estimate a causal network of 29 cardiovascular-related proteins and linked the estimated network to Alzheimer's disease (AD). A simulation study was conducted to demonstrate the effectiveness of this method. An R package implementing the proposed method and an R Shiny App for the visualization of the estimated protein network are being made available.

Keywords: Alzheimer's disease (AD); Directed acyclic graph (DAG); Genome-wide association study (GWAS); Likelihood ratio test; Proteomics.

*Correspondence to W. Pan (panxx014@umn.edu) and T. Yang (yang3704@umn.edu). [1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A. [2]School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.

1 Introduction

Network analysis has deepened our understanding of biological mechanisms and disease etiologies [1]. Specifically, protein-protein interaction (PPI) networks that capture interplay of proteins in the biomolecular systems are vital for normal cell functions [2]. Disturbing of the normal pattern in the PPI network can be causative to or indicative of a disease state. Studies have linked co-regulatory networks of proteins to a variety of complex diseases [3, 4]. Recently, a network-based method modeling PPI boasted high accuracy rates in cancer prediction [5]. Cheng et al. [6] further showed that disease-associated variants were significantly enriched in the sequences coding PPI interfaces compared to variants in healthy individuals. Their work also demonstrated associations of PPIs with drug resistance and overall survival, highlighting the use of protein networks for informing genotype-based therapy. Network-based analyses have shown their potential in advancing precision medicine for complex diseases over traditional approaches which focus on monogenic mutations and independent assessment of risk factors [7].

Network analyses can be categorized into two groups. One group of methods utilizes only phenotypic data to construct networks. For example, weighted gene network co-expression analysis estimates an undirected network which is further characterized using dimension reduction techniques [8]. Graphical lasso formulation employs penalized methods to estimate a Gaussian graphical model for a large number of variables[9]. Bayesian network analysis, another widely used approach, also only requires phenotypic data to estimate a directed acyclic graph (DAG) [10], and recent improvements in computational approaches have led to much shortened computational time [11]. The other group of methods exploits the use of instrumental variable (IV) techniques to estimate a DAG, assuming a linear structural equation model where genetic variants serve as IVs for the phenotypic variables. Chen et al. [12] developed a penalized two-stage least squares method to estimate a DAG, assuming known intervention targets. Li et al. [13] further extended the work to accommodate unknown intervention targets commonly encountered in biological applications.

Individual-level data are required for all the methods above, which, however, can be

difficult to obtain, especially for human studies, due to logistic limitations and privacy concerns. On the other hand, many genome-wide association studies (GWAS) have shared their summary statistics publicly, generating a rich and valuable data resource. Thus, we propose adapting the network estimation and inference methods of Li et al. [13] to rely only on GWAS summary statistics and a genetic reference panel. We will show how a DAG can be estimated for cardiovascular-related proteins using a large-scale proteomic GWAS summary dataset, then link the protein network to Alzheimer's disease (AD). Our algorithm for the proposed work is packaged in R, which will be available soon.

2 Methods

2.1 Network modeling and data

2.1.1 Directed phenotype network

Our goal is to use genotypes as external interventions to construct and infer a DAG that describes the directed relationships among a set of phenotypes. Biological graphs such as these in Gene Ontology [14] are usually represented in the form of DAG. In the framework of interventional Gaussian DAG (IntDAG) [13], we assume

$$\boldsymbol{Y} = \boldsymbol{Y}\mathbf{U} + \boldsymbol{X}\mathbf{W} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P)$ is the $N \times P$ data matrix of P phenotypes, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_Q)$ is the $N \times Q$ data matrix of Q genotypes serving as IVs, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_P)$ is the $N \times P$ error matrix with each row sampled from $N(\mathbf{0}, \mathrm{Diag}(\omega_1^2, \ldots, \omega_P^2))$, and N is the sample size.

In Equation (1), $\mathbf{U}$ and $\mathbf{W}$ are unknown parameters to be estimated. The $P \times P$ matrix $\mathbf{U} = (\mathrm{u}_{kj})_{1 \le k,j \le P}$ specifies the network structure such that $\mathrm{u}_{kj} \neq 0$ indicates a directed relation from phenotype k to phenotype j. The $Q \times P$ matrix $\mathbf{W} = (\mathrm{w}_{qp})_{1 \le q \le Q}^{1 \le p \le P}$ specifies the targets and strengths of interventions in that $\mathrm{w}_{qp} \neq 0$ indicates an interventional relation from genotype q to phenotype p. Let $\mathcal{E} = \{(k, j) : \mathrm{u}_{kj} \neq 0\}$ be the set of directed relations and $\mathcal{I} = \{(q, p) : \mathrm{w}_{qp} \neq 0\}$ be the set of interventional relations.

Without loss of generality, we assume each phenotype and all the genotype are centered at mean 0.

2.1.2 Summary statistics and reference panel

In Equation (1), the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ contain individual information such that each row represents the variables measured on an individual. On the other hand, GWAS summary statistics aggregate N observations into a single measure for each single nucleotide polymorphism (SNP). This measure is the average effect of having one copy of the effect allele on the phenotype being studied. It is estimated by $\hat{\beta}_{qp} \approx (\boldsymbol{x}_q^\top \boldsymbol{x}_q)^{-1} \boldsymbol{x}_q^\top \boldsymbol{y}_p$, often reported along with accompanying statistics such as the corresponding standard error (SE($\hat{\beta}_{qp}$)), z-score (z_{qp}), sample size (N), reference allele (REF), minor allele frequency (MAF) and p-value, as shown below,

$$\begin{array}{ccccccc} \hat{\beta}_{qp} & \text{SE}(\hat{\beta}_{qp}) & z_{qp} & \text{p-value} & \text{REF} & \text{MAF} & N \\ \end{array}$$
$$\begin{pmatrix} 0.13 & 0.01 & 13.0 & 1.02\times10^{-38} & A & 0.45 & 14{,}000 \\ 0.20 & 0.06 & 3.3 & 0.0004 & G & 0.05 & 14{,}000 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.11 & 0.05 & 2.20 & 0.0228 & T & 0.32 & 14{,}000 \end{pmatrix}.$$

The summary statistics contain the Q SNPs in $\mathbf{W}$. As a complement to the summary-level data, a reference panel [15, 16, 17, 18], comprising genotypic data of individuals from a general population, provides the correlation structure among the genotypes. Given an $N_r \times Q$ (standardized) reference panel $\boldsymbol{X}_r$ of N_r individuals, we follow the conventional suggestion [19] to regularize the correlation matrix $\boldsymbol{R}$ such that

$$\boldsymbol{R} = (1 - s_p)\boldsymbol{X}_r^\top \boldsymbol{X}_r + s_p \boldsymbol{I},$$

where $0 \leq s_p \leq 1$ is a real number controlling the degree of regularization.

From the summary statistics and the reference panel, we compute the following quantities that are used for the construction and inference of the directed phenotype network.

- The covariance matrix of genotypes $\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$ is estimated by $\frac{1}{N}\widehat{\boldsymbol{X}^\top\boldsymbol{X}} = \frac{1}{N_r}\boldsymbol{R}$.

- Let $s_q^2 = \frac{1}{N}\boldsymbol{x}_q^\top\boldsymbol{x}_q$. Then s_q^2 is estimated by $\hat{s}_q^2 = 2 \times \text{MAF} \times (1 - \text{MAF})$ provided that MAF is reported in the summary statistics, or otherwise estimated by the qth diagonal element of $\frac{1}{N_r}\boldsymbol{X}_r^\top\boldsymbol{X}_r$, $\hat{s}_q^2 = \frac{1}{N_r}(\boldsymbol{X}_r^\top\boldsymbol{X}_r)_{qq}$.

- Given $\hat{s}_q^2$, $\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{y}_p$ is estimated by $\frac{1}{N}\widehat{\boldsymbol{X}^\top\boldsymbol{y}_p} = \left(\hat{s}_1^2\hat{\beta}_{p1}, \hat{s}_2^2\hat{\beta}_{p2}, \ldots, \hat{s}_Q^2\hat{\beta}_{pQ}\right)^\top$.

- For $\frac{1}{N}\boldsymbol{y}_p^\top\boldsymbol{y}_p$, we use the median estimate $\frac{1}{N}\widehat{\boldsymbol{y}_p^\top\boldsymbol{y}_p} = \underset{1\le q\le Q}{\text{med}}\left(N \times \hat{s}_q^2 \times \text{SE}(\hat{\beta}_{pq})^2 + \hat{s}_q^2 \times \hat{\beta}_{pq}^2\right)$.

- Finally, $\frac{1}{N}\boldsymbol{y}_k^\top\boldsymbol{y}_i$ is estimated using the null SNPs from GWAS summary statistics, i.e., SNPs not marginally associated with $\boldsymbol{y}_k$ or $\boldsymbol{y}_i$. Following [20], $\text{Corr}(\boldsymbol{y}_k, \boldsymbol{y}_i) \approx \text{Corr}(\boldsymbol{z}_k, \boldsymbol{z}_i)$ where $\boldsymbol{z}_k$ and $\boldsymbol{z}_i$ are vectors of z-scores for the null SNPs. Thus, we can rearrange the sample correlation formula for (centered) phenotype variables and plug in our approximation to obtain $\frac{1}{N}\widehat{\boldsymbol{y}_k^\top\boldsymbol{y}_i} = \frac{1}{N}\,\text{Corr}(\boldsymbol{z}_k, \boldsymbol{z}_i)\sqrt{\widehat{\boldsymbol{y}_k^\top\boldsymbol{y}_k}\widehat{\boldsymbol{y}_i^\top\boldsymbol{y}_i}}$. In practice, the SNPs with p-values larger than 0.05 are considered as null SNPs. An alternative method to consider for estimating $\frac{1}{N}\boldsymbol{y}_k^\top\boldsymbol{y}_i$ is [21], which is also feasible to use for GWAS although not used herein.

Next we extend the framework of intDAG to leverage large-scale GWAS summary statistics.

2.2 Method for network construction

The construction of intDAG consists of three steps.

1. First, we use penalized regressions to estimate the genotype-phenotype association matrix $\mathbf{V} := \mathbf{W}(\mathbf{I} - \mathbf{U})^{-1}$ in the following equation

$$\boldsymbol{Y} = \boldsymbol{X}\mathbf{V} + \boldsymbol{\epsilon}(\mathbf{I} - \mathbf{U})^{-1}. \tag{2}$$

2. Next, we employ the peeling algorithm [13] to learn a super-DAG (a directed super-graph without cycles) based on $\hat{\mathbf{V}}$ obtained in Step 1.

3. Finally, we estimate $\mathbf{U}$ and $\mathbf{W}$ through penalized regressions based on the estimated super-DAG in Step 2.

Now, we elaborate on our extensions to accommodate summary statistics.

2.2.1 Estimation of V by truncated Lasso penalized regressions

In Equation (2), the matrix $\mathbf{V}$ can be estimated column-wise from

$$\boldsymbol{y}_p = \boldsymbol{X}\mathbf{v}_p + \boldsymbol{\xi}_p, \quad \boldsymbol{\xi}_p \sim N(\mathbf{0}, \sigma_p^2\mathbf{I}), \tag{3}$$

where $\boldsymbol{y}_p$ is the data vector of phenotype p, $\boldsymbol{X}$ is the data matrix of genotypes, and vector $\mathbf{v}_p = (\mathrm{v}_{1p}, \ldots, \mathrm{v}_{Qp})^\top$ is the pth column of $\mathbf{V}$. Given the summary statistics, we expand the squared error function $\|\boldsymbol{y}_p - \boldsymbol{X}\mathbf{v}_p\|^2 = \boldsymbol{y}_p^\top\boldsymbol{y}_p + \mathbf{v}_p^\top\boldsymbol{X}^\top\boldsymbol{X}\mathbf{v}_p - 2\boldsymbol{y}_p^\top\boldsymbol{X}\mathbf{v}_p$, and replace the quantities $\boldsymbol{y}_p^\top\boldsymbol{y}_p$, $\boldsymbol{X}^\top\boldsymbol{y}_p$, and $\boldsymbol{X}^\top\boldsymbol{X}$ in Equation (3) with their estimates $\widehat{\boldsymbol{y}_p^\top\boldsymbol{y}_p}$, $\widehat{\boldsymbol{X}^\top\boldsymbol{y}_p}$, and $\widehat{\boldsymbol{X}^\top\boldsymbol{X}}$, respectively. As a result, we estimate $\mathbf{v}_p$ through regressions with the Truncated Lasso Penalty (TLP) [22] to minimize

$$\mathbf{v}_p^\top\widehat{\boldsymbol{X}^\top\boldsymbol{X}}\mathbf{v}_p - 2\mathbf{v}_p^\top\widehat{\boldsymbol{X}^\top\boldsymbol{y}_p} \quad \text{s.t.} \quad J(\mathbf{v}_p, \tau_p) \le \kappa_p, \tag{4}$$

where $\kappa_p > 0$ is an integer tuning parameter and $J(\mathbf{v}_p, \tau_p) = \sum_{q=1}^{Q}\min(|\mathrm{v}_{qp}|/\tau_p, 1)$ is the TLP function, which does not penalize the parameters over the threshold τ_p. We use the R package "glmtlp" [23] to fit the summary-level data regression (4).

For implementation, we fix $\tau_p = 0.5\sqrt{\log(Q)/N}$ and choose $\kappa_p \in \{1, \ldots, Q\}$ individually for each of the P penalized regressions by minimizing the pseudo-BIC [24] defined as

$$\mathrm{pBIC}(\lambda_p, \tau_p) = \widehat{\mathrm{SSE}}_p(\hat{\mathbf{v}}_p)/\hat{\sigma}_p^2 + \|\hat{\mathbf{v}}_p\|_0 \times \log(N),$$

where $\widehat{\mathrm{SSE}}_p(\hat{\mathbf{v}}_p) = \widehat{\boldsymbol{y}_p^\top\boldsymbol{y}_p} - 2\hat{\mathbf{v}}_p^\top\widehat{\boldsymbol{X}^\top\boldsymbol{y}_p} + \hat{\mathbf{v}}_p^\top\widehat{\boldsymbol{X}^\top\boldsymbol{X}}\hat{\mathbf{v}}_p$ is the (estimated) sum of squared error of $\hat{\mathbf{v}}_p$, $\|\hat{\mathbf{v}}_p\|_0$ is the number of nonzero coefficients in $\hat{\mathbf{v}}_p$, $\hat{\mathbf{v}}_p$ is the estimate in (4) with tuning parameters (λ_p, τ_p), N is the sample size (when N differs, the median is taken), and $\hat{\sigma}_p^2$ is the estimated residual variance for phenotype p in Equation (3). When Q is small compared

to N as in our application, a consistent estimate for σ_p^2 can be obtained from the ordinary least squares using all Q genotypes,

$$\hat{\sigma}_p^2 = (\widehat{\boldsymbol{y}_p^\top \boldsymbol{y}_p} - 2\hat{\mathbf{v}}_{p,olse}^\top \widehat{\boldsymbol{X}^\top \boldsymbol{y}_p} + \hat{\mathbf{v}}_{p,olse}^\top \widehat{\boldsymbol{X}^\top \boldsymbol{X}} \hat{\mathbf{v}}_{p,olse})/(N-Q),$$

where $\hat{\mathbf{v}}_{p,olse}$ is the estimate in Equation (4) with $\kappa_p = P$.

Letting $\hat{\mathbf{v}}_p$ $(1 \leq p \leq P)$ be the estimates with the optimally chosen tuning parameters, the final estimate of $\mathbf{V}$ is $\hat{\mathbf{V}} = (\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_P)$.

2.2.2 Estimation of super-DAG by the peeling algorithm

Given $\hat{\mathbf{V}}$, the peeling algorithm [13] can be used to construct a super-DAG with phenotype edge set $\mathcal{E}^+$ (a superset of $\mathcal{E}$) and interventional edge set $\mathcal{I}^+$ (a superset of $\mathcal{I}$). The key idea is that the sparse pattern of matrix $\mathbf{V}$ characterizes the orientations of the relations among the phenotypes. Specifically, it is demonstrated in [13] that $\mathrm{v}_{qp} \neq 0$ implies that genotype q intervenes on phenotype p or an ancestor node of phenotype p in the DAG. Thus, if $\mathrm{v}_{qp} \neq 0$ and $\mathrm{v}_{qi} = 0$ for $i \neq p$, then phenotype p is a leaf node in the DAG, that is, there is no directed edge from phenotype p to the others. On this basis, we can sequentially identify and remove (i.e., peel) the leaf node in the DAG, and construct supersets $\mathcal{E}^+$ and $\mathcal{I}^+$. Since the peeling algorithm solely depends on $\hat{\mathbf{V}}$, no modification is needed for summary-level data.

2.2.3 Estimation of U and W

The peeling algorithm yields supersets $\mathcal{E}^+ \supseteq \mathcal{E}$ and $\mathcal{I}^+ \supseteq \mathcal{I}$. To remove the extra edges in $\mathcal{E}^+$ and $\mathcal{I}^+$, we consider fitting $\mathbf{U}$ and $\mathbf{W}$ within a restricted model defined by $\mathcal{E}^+$ and $\mathcal{I}^+$. From Equation (1), for phenotype p, we have

$$\boldsymbol{y}_p = \sum_{k \in \mathcal{E}^+(p)} \boldsymbol{y}_k \mathrm{u}_{kp} + \sum_{q \in \mathcal{I}^+(p)} \boldsymbol{x}_q \mathrm{w}_{qp} + \boldsymbol{\epsilon}_p, \quad \boldsymbol{\epsilon}_p \sim N(\mathbf{0}, \omega_p^2 \mathbf{I}), \tag{5}$$

where $\mathcal{E}^+(p) = \{k : (k,p) \in \mathcal{E}^+\}$ and $\mathcal{I}^+(p) = \{q : (q,p) \in \mathcal{I}^+\}$. As in Section 2.2.1, we replace the corresponding quantities with the summary-level data estimates, and fit the TLP

regression

$$\min_{\boldsymbol{\gamma}} \ \boldsymbol{\gamma}^\top \begin{pmatrix} (\widehat{\boldsymbol{Y}^\top \boldsymbol{Y}})_{\mathcal{E}^+(p),\mathcal{E}^+(p)} & (\widehat{\boldsymbol{Y}^\top \boldsymbol{X}})_{\mathcal{E}^+(p),\mathcal{I}^+(p)} \\ (\widehat{\boldsymbol{X}^\top \boldsymbol{Y}})_{\mathcal{I}^+(p),\mathcal{E}^+(p)} & (\widehat{\boldsymbol{X}^\top \boldsymbol{X}})_{\mathcal{I}^+(p),\mathcal{I}^+(p)} \end{pmatrix} \boldsymbol{\gamma} - 2\boldsymbol{\gamma}^\top \begin{pmatrix} (\widehat{\boldsymbol{Y}^\top \boldsymbol{Y}})_{\mathcal{E}^+(p),p} \\ (\widehat{\boldsymbol{X}^\top \boldsymbol{Y}})_{\mathcal{I}^+(p),p} \end{pmatrix} \quad \text{s.t.} \quad J(\boldsymbol{\gamma}, \tau_p) \leq \kappa_p \tag{6}$$

where $\boldsymbol{\gamma} = ((\mathrm{u}_{kp})_{k\in\mathcal{E}^+(p)}, (\mathrm{w}_{kp})_{k\in\mathcal{I}^+(p)})$ is the parameter vector and $J(\boldsymbol{\gamma}, \tau_p) = \sum_k \min(|\gamma_k|/\tau_p, 1)$. We fix $\tau_p = 0.5\sqrt{\log(|\mathcal{E}^+(p)| + |\mathcal{I}^+(p)|)/N}$ and the tuning parameter $\kappa_p \in \{1, \ldots, |\mathcal{E}^+(p)| + |\mathcal{I}^+(p)|\}$ are selected by pseudo-BIC as described in Section 2.2.1. The estimated $\hat{\mathbf{u}}_p = ((\hat{\mathrm{u}}_{kp})_{k\in\mathcal{E}^+(p)}, \mathbf{0})$ and $\hat{\mathbf{w}}_p = ((\hat{\mathrm{w}}_{qp})_{q\in\mathcal{I}^+(p)}, \mathbf{0})$ $(1 \leq p \leq P)$ are aggregated to form the the final estimate $\hat{\mathbf{U}} = (\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_P)$ and $\hat{\mathbf{W}} = (\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_P)$.

Due to penalization, we recommend following the common practice to standardize the variables so that the phenotypes and genotypes are on a comparable scale. Moreover, if only $\mathbf{U}$ is of interest, penalization of $\mathbf{W}$ is optional.

2.3 Likelihood-based inference for a DAG

We extend the likelihood ratio inference [13] to quantify the uncertainty of the network structures. As in [13], we consider two types of hypothesis testing.

- **Testing of multiple directed relations.** The null hypothesis $H_0 : \mathrm{u}_{kj} = 0$ for each $(k, j) \in \mathcal{H}$ and alternative hypothesis $H_a : \mathrm{u}_{kj} \neq 0$ for some $(k, j) \in \mathcal{H}$. Rejecting H_0 indicates evidence for the presence of some hypothesized relationships in the network.

- **Testing of a directed pathway.** The null hypothesis $H_0 : \mathrm{u}_{kj} = 0$ for some $(k, j) \in \mathcal{H}$ and alternative hypothesis $H_a : \mathrm{u}_{kj} \neq 0$ for each $(k, j) \in \mathcal{H}$. Rejecting H_0 indicates evidence for the presence of the entire directed pathway in the phenotype network.

The procedure for testing multiple directed relations comprises five steps.

1. Estimate $\mathbf{V}$ and use the peeling algorithm to obtain $\mathcal{E}^+$ and $\mathcal{I}^+$ as in Section 2.2.

2. Identify the set of nondegenerate edges $\mathcal{D}$ [13], which contains $\mathcal{D}_p$, the nondegenerate edges pointed to phenotype p.

3. Estimate the parameters $\mathbf{U}$ and $\mathbf{W}$ under H_0 and H_a, respectively. Specifically, denote by $\hat{\mathbf{U}}^{(0)}$ and $\hat{\mathbf{W}}^{(0)}$ the estimates under H_0. Then $\hat{\mathbf{U}}^{(0)}$ and $\hat{\mathbf{W}}^{(0)}$ are computed as in the regression (6) with an additional constraint that $\mathrm{u}_{kj} = 0$ for $(k, j) \in \mathcal{H}$. Let $\hat{\mathbf{U}}^{(1)}$ and $\hat{\mathbf{W}}^{(1)}$ be the estimates under H_a. Then $\hat{\mathbf{U}}^{(1)}$ and $\hat{\mathbf{W}}^{(1)}$ are computed from the restricted models $(1 \leq p \leq P)$,

$$\boldsymbol{y}_p = \sum_{k \in \mathcal{E}^+(p) \cup \mathcal{D}_p} \boldsymbol{y}_k \mathrm{u}_{kp} + \sum_{q \in \mathcal{I}^+(p)} \boldsymbol{x}_q \mathrm{w}_{qp} + \boldsymbol{\epsilon}_p, \quad \boldsymbol{\epsilon}_p \sim N(\mathbf{0}, \omega_p^2 \mathbf{I}),$$

via regression (6), where the penalties become $J(\mathbf{u}_p, \tau_p) = \sum_{k \in \mathcal{E}^+(p) \setminus \mathcal{D}_p} \min(|\mathrm{u}_{kp}|, \tau_p)$ and $J(\mathbf{w}_p, \tau_p) = \sum_{q \in \mathcal{I}^+(p)} \min(|\mathrm{w}_{qp}|, \tau_p)$.

4. Compute $\hat{\omega}_p^2$ $(1 \leq p \leq P)$ from the residual sum of squares of $(\hat{\mathbf{u}}_p^{(1)}, \hat{\mathbf{w}}_p^{(1)})$.

5. Compute the test statistic as

$$T = 2L\Big(\hat{\mathbf{U}}^{(1)}, \hat{\mathbf{W}}^{(1)}, \hat{\omega}_1^2, \ldots, \hat{\omega}_P^2\Big) - 2L\Big(\hat{\mathbf{U}}^{(0)}, \hat{\mathbf{W}}^{(0)}, \hat{\omega}_1^2, \ldots, \hat{\omega}_P^2\Big), \tag{7}$$

where L is the log-likelihood of the model (Equation (1)). By [13], T is approximately chi-squared distributed with degrees of freedom $|\mathcal{D}|$ when the size $|\mathcal{D}|$ is less than 50; $(T - |\mathcal{D}|)/\sqrt{2|\mathcal{D}|}$ is approximately standard normal when $|\mathcal{D}|$ exceeds 50. The p-value is calculated accordingly.

The procedure for testing a directed pathway is similar, with minor modifications.

1'. Estimate $\mathbf{V}$ as in Step 1.

2'. First, we decompose H_0 into each nongenerate edge $\left\{H_0^{[(k,j)]} : \mathrm{u}_{(k,j)} = 0\right\}_{(k,j) \in \mathcal{H}}$. For each $H_0^{[(k,j)]}$, implement Steps 2–5 above to obtain the corresponding p-values. The final p-value is computed as the maximum of the p-values for the sub-hypotheses.

2.4 A real data application on protein network

The role of cardiovascular diseases has been recognized as an important etiologic hallmark of AD [25]. There are different hypotheses on the various mechanisms underlying the asso-

ciation between AD and cardiovascular diseases [26]. In this real data application, we constructed a directed PPI network of some cardiovascular-related proteins based on a GWAS of 83 plasma protein biomarkers. We further connected the PPI network to AD through MR analyses.

2.4.1 GWAS summary statistics for cardiovascular-related proteins

The GWAS summary statistics on 83 cardiovascular-related proteins which came from Wald tests for the association between each SNP and the standardized residuals among 3394 individuals by Folkersen et al. [27] were used. Five proteins were removed from the analysis as their corresponding protein-encoding genes are located on the sex chromosome. The summary statistics were processed to remove indels, SNPs located within one base pair of an indel, and SNPs with INFO ≤ 0.3 or MAF ≤ 0.05. We used the following steps to select putative IVs for the proteins: 1) SNPs were clumped at an r^2 value of 0.01 to ensure SNPs were independent of each other; 2) Only the SNPs in the clumped data files located within $\pm$1MB of each protein-encoding genes were considered. In general, cis-regulatory changes will be less pleiotropic [28] and thus these SNPs located close to the genes are more likely to be valid IVs due to the exclusion assumption [29, 30, 13]; 3) To ensure the relevance assumption [13] was satisfied, we only selected SNPs whose p-values were below the GWAS significance threshold (5×10^{-8}). This filtering process led to a total number of 57 SNPs and 29 proteins with at least one putative IV in the final network analysis. The empirical correlation of each pair of proteins was calculated as the correlation matrix of the z-scores of all of the autosomal SNPs with MAF ≥ 0.05, INFO ≥ 0.3, and GWAS p-values ≥ 0.05 for both proteins.

2.4.2 GWAS summary statistics for for AD

We explored the relationship between each of the 29 proteins in Folkersen et al. [27] and AD. We used the summary statistics of the GWAS for AD from a most recent study totaling 111,326 clinically diagnosed or "proxy" AD cases and 677,663 controls [31]. We removed

SNPs with MAF < 0.05, SNPs not included in the GWAS of Folkersen et al.[27], and clumped SNPs for LD at $r^2 = 0.01$. Among the remaining SNPs, we selected IVs only with GWAS p-value $< 5 \times 10^{-8}$ in the MR analyses.

2.5 Simulation studies

We simulated the data assuming a fixed $\mathbf{U}$, $\mathbf{W}$, standardized genotype matrix $\boldsymbol{X}$, and sampled each row of $\boldsymbol{\epsilon}$ independently from $N(\mathbf{0}, \mathrm{Diag}(\omega_1^2, \ldots, \omega_P^2))$. Then we generated $\boldsymbol{Y}$ from equation:

$$\boldsymbol{Y} = \mathbf{V}^\top \boldsymbol{X} + \boldsymbol{\epsilon}(\mathbf{I} - \mathbf{U})^{-1}.$$

Without loss of generality, no intercept was modeled, i.e., $\boldsymbol{Y}$ was centered at mean 0. The values of $\mathbf{U}$, $\omega_1^2, \ldots, \omega_P^2$, and $\mathbf{W}$ are provided in Supplemental Material S1.1, where the structure of the relationship of $\boldsymbol{Y}$ followed a DAG of 15 nodes/phenotypes (Figure 1). The effect sizes of the non-zero components of $\mathbf{U}$ ranged from 0.002 to 1.16 with a median of 0.06. All phenotypes had at least one valid IV. Twenty-six SNPs were included in the model, with their effect sizes ranged from -2.2 to 2.5 with a median of -0.11. Two SNPs violated the relevance assumption, while the rest were valid IVs. We also varied the effect sizes to be 1/3 and 1/15 of $\mathbf{U}$ while keeping $\mathbf{W}$ fixed.

The standardized genotype matrix $\boldsymbol{X}$ was obtained from unrelated individuals of European ancestry in UK Biobank [18] (i.e., individuals with kinship coefficients larger than 0.084 were removed). We then calculated the summary statistics using a linear model of each phenotype on teach standardized genotype, and input the summary statistics into the proposed algorithm. The reference panel was obtained from the same population and was independent of $\boldsymbol{X}$. SNPs on chromosome 22 with a Hardy-Weinberg Disequilibrium test p-value > 0.0001, missing call rate < 0.05, and MAF > 0.05 were pruned to have $r^2 < 0.01$. We then randomly selected 26 SNPs for $\boldsymbol{X}$. Missing values of SNPs were imputed by their mean. Null SNPs were directly simulated to be independent of each other and had no relationship with $\boldsymbol{Y}$.

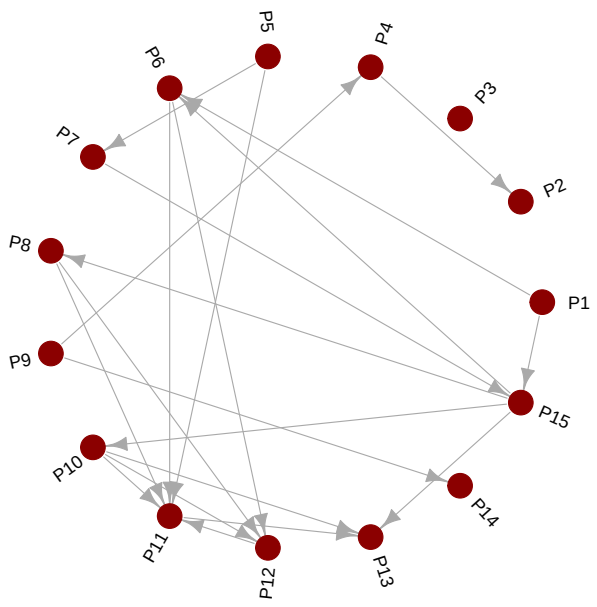

Figure 1: True DAG for the simulation study with 15 phenotypes

We evaluated the performance of both the network construction and statistical inference for the proposed method. To evaluate the performance of network construction, we examined the false positive (TP) and false negative (FN) rates for $\hat{\mathbf{U}}$ over 200 replications. The sample size of the summary statistics was varied at 3000, 6000, 9000, and 12,000, and the sample size of the reference panel was fixed at 3000. Null SNPs were simulated to have the same sample size as the GWAS summary statistics.

In terms of testing, we examined the empirical Type I error/power of the likelihood ratio tests with increasing sample sizes and varying strength of $\mathbf{U}$ for the following five scenarios over 1000 replications for the two types of testing.

I. Testing one or more directed relations:

- A1. Type 1 error: testing one edge when in truth it was null with $H_0 : \mathrm{u}_{1,14} = 0$ vs $H_1 : \mathrm{u}_{1,14} \neq 0$;

- A2. Power: testing one edge when in truth it was not null with $H_0 : \mathrm{u}_{1,6} = 0$ vs $H_1 : \mathrm{u}_{1,6} \neq 0$;

- A3. Power: testing two edges together when in truth both were not null with $H_0 : \mathrm{u}_{7,15} = \mathrm{u}_{1,6} = 0$ vs $H_1 : \mathrm{u}_{7,15} \neq 0$ or $\mathrm{u}_{1,6} \neq 0$;

II. Testing of a directed pathway:

- B1. Type 1 error: testing whether at least one edge was not null when in truth only one was not null with $H_0 : \mathrm{u}_{1,14} = 0$ or $\mathrm{u}_{6,12} = 0$ vs $H_1 : \mathrm{u}_{1,14} \neq 0$ and $\mathrm{u}_{6,12} \neq 0$;

- B2. Power: testing whether at least one edge was not null when in truth both were not null with $H_0 : \mathrm{u}_{1,6} = 0$ or $\mathrm{u}_{6,12} = 0$ vs $H_1 : \mathrm{u}_{1,6} \neq 0$ and $\mathrm{u}_{6,12} \neq 0$;

The true strengths of the tested edges were: $\mathrm{u}_{1,14} = 0$, $\mathrm{u}_{1,6} = 0.27$, $\mathrm{u}_{7,15} = 0.44$, and $\mathrm{u}_{6,12} = 1.36$.

3 Results

3.1 Real data application results

We constructed a DAG of the 29 proteins as described in Section 2.5.1 and performed MR analysis on each protein to evaluate their relationship with AD using the TwoSampleMR package. Among all proteins, RETN did not have any IV, and thus was excluded from the MR analysis. We used Egger's test of intercept for examining the exclusion assumption: If a protein had a p-value of the Egger's test of intercept $> 0.05/29$, there was no evidence against no direct/pleiotropic effects, and we'd go with the more powerful MR-IVW method; otherwise, we used MR-Egger (to allow pleiotropic effects of IVs). In any case, we used the p-value cut-off $< 0.05/29$ to declare statistical significance. Table S1 contains a complete list of MR results. The protein IL18, which showed marginal significance in both Egger's test of intercept (p-value = 0.06) and MR-Egger (p-value = 0.07), was a parent node for several proteins related to AD, including CHI3L1, CTSD, CXCL16, HAVCR1, and NPPB.

We further performed the likelihood ratio test on each edge. Edges with p-value $< 0.05/(29 \times 28 - 66)$ were considered as significant and were in solid line in Figure 2. The number of tests in the Bonferonni correction is bounded by the sum of possible edges among all the nodes minus the total number of edges after the peeling algorithm, which is justified in Supplementary Material S2.1. Each edge from IL18 to the five AD-associated proteins was highly significant in the likelihood ratio test, thus suggesting that a simultaneous testing of the pathway from IL18 to the five genes would be significant. Previous studies detected an increased levels of pro-inflammatory IL18 in both cardiovascular diseases and in brain regions of AD patients [32]. IL18 is known to increase the level of Cdk5 and GSK-3β, which are involved in Tau hyperphosphorylation, and the inhibition of Cdk5 was known to improve AD subject's condition [33]. Our work suggests a possible regulatory role of IL18 on multiple AD-associated proteins. According to OpenTargets.org [34] for current pharmaceuticals either approved or in development with IL18, this protein is currently a target of an antibody drug to treat diabetes mellitus and a few other conditions; diabetes has long been linked to AD with epidemiological and biological evidences [35].

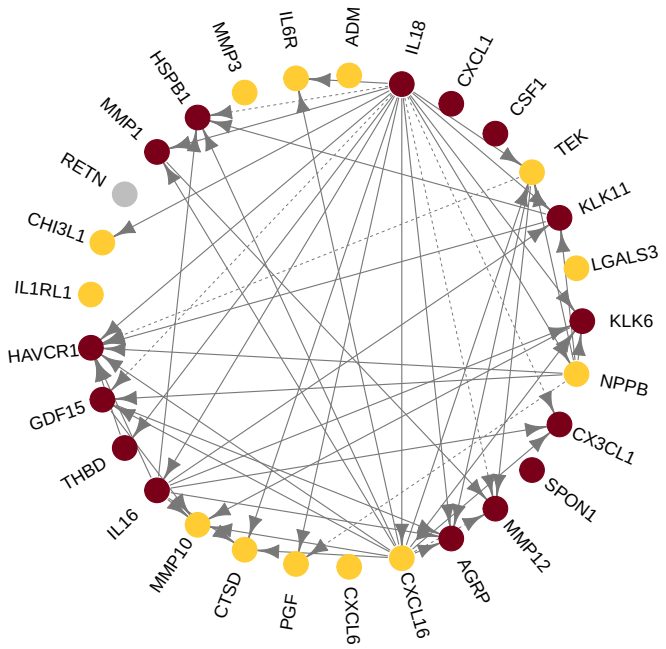

Figure 2: Estimated DAG for 29 proteins based on the GWAS summary statistics of Folkersen et al [27]; Proteins significantly associated with AD in MR analysis are colored gold. A solid line represents an edge that is statistically significant by the likelihood ratio test whereas a dash line represents an edge that is not significant.

3.2 Simulation results

With the increase of sample size, the constructed networks became more closely to the true graph (Figure 3, Numeric values in Figure 3.A are in Table S2 - S5). More specifically, FP of $\widehat{\mathbf{U}}$ was around 0.05 and FN of $\widehat{\mathbf{U}}$ decreased with the increase of sample size when 15000 null SNPs were used to estimate the 15×15 matrix of $\boldsymbol{Y}^\top \boldsymbol{Y}$. Unsurprisingly, the performance of using $\widehat{\boldsymbol{Y}^\top \boldsymbol{Y}}$ estimated by 15,000 null SNPs (denoted as $\mathbf{U}'$ in Figure 3) was slightly worse than that of using $\boldsymbol{Y}^\top \boldsymbol{Y}$ (denoted as $\mathbf{U}$ in Figure 3). In addition, the FP of $\widehat{\mathbf{U}}$ clearly decreased with the increase of effect sizes (from $\mathbf{U}/15$ to $\mathbf{U}/3$ to $\mathbf{U}$). To check the validity of our method, we compared pBIC estimated from summary statistics with BIC estimated

from the individual-level data for the same set of penalized regression coefficients. We found the two sets of values were highly concordant and the results from one iteration was plotted in Figure S2.

In terms of testing, we observed well-controlled Type I error rates for scenarios A1 and B1 (Numeric values present in Figure 3, B-D are in Table S6 - S9), using $\boldsymbol{Y}^\top\boldsymbol{Y}$. We note that the empirical Type I error rates might become conservative when $\boldsymbol{Y}^\top\boldsymbol{Y}$ is replaced by its estimate $\widehat{\boldsymbol{Y}^\top\boldsymbol{Y}}$ derived from 15000 null SNPs (Table S9). In real data analysis of GWAS summary level data, typically a much larger number of null SNPs and various other strategies can be used to better estimate $\boldsymbol{Y}^\top\boldsymbol{Y}$ [21, 20, 36]; however, the investigation in this direction is out of our scope. Furthermore, empirical power was high for scenarios A2, A3, and B2 and increased with sample size and effect size. The empirically power of jointly testing two edges $\mathrm{u}_{1,6}$ and $\mathrm{u}_{7,15}$ (scenario A3) was larger than testing one edge $\mathrm{u}_{1,6}$ alone (scenario A2).

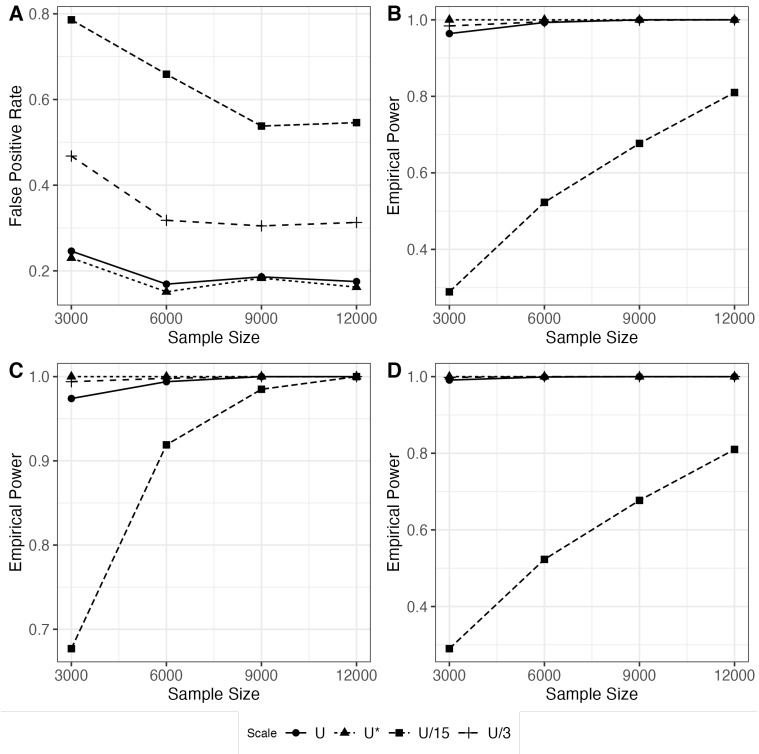

Figure 3: Performance of network construction (A) and likelihood ratio tests (B-D) in simulation with varying sample sizes and true effect sizes at **U**, **U**/3, and **U**/15. Figures B, C, and D represent the scenario A2, A3, and B2 respectively.

4 Discussion

In this paper, we present a method to estimate an IntDAG of phenotypes utilizing linear structural equation models, applicable to GWAS summary statistics in the absence of individual-level data. We demonstrated a satisfactory performance in terms of the FP and FN rates in network construction and high empirical power and of well-controlled Type I error rates of the likelihood ratio tests. We applied this method to a large-scale proteomic GWAS summary dataset to obtain an estimated DAG of 29 cardiovascular-related proteins, and further illustrated the effects of these proteins on AD by MR analysis. These results can be useful in understanding the disease etiology, drug repurposing, and other applications for AD. We note that in the real data application, we used uncorrelated SNPs as IVs. However,

the proposed IntDAG model allows IVs to be correlated. As the putative IVs were abundant, using uncorrelated IVs simplified the estimation of $\boldsymbol{X}^\top \boldsymbol{X}$ (i.e., it became a diagonal matrix), such that the potential problems with the use of a reference panel can be avoided [37].

In recent years, a large amount of summary-level data becomes widely accessible. On the molecular level, many studies published their summary statistics for SNP-molecular phenotype associations. For example, variant-gene associations in 49 tissues can be directly downloaded from GTExPortal [38]. Beyond the molecular phenotypes, UK Biobank alone provides GWAS summary statistics on more than 7000 traits, including but not limited to cognitive functions, early life factors, health and medical history, and physical measurement [18]. Our proposed method provides a computational and analytical tool to explore the relationships among multiple phenotype variables by taking advantage of rapid advances in GWAS and other association mappings.

4.1 Software

All preparation of the reference panel and GWAS data for both the DAG estimation and MR analysis was done using PLINK version 1.9 [39]. All the analyses were performed in R where all MR analyses were performed using the TwoSampleMR package [40].

References

[1] Peng Zhang and Yuval Itan. Biological network approaches and applications in rare disease studies. *Genes*, 10(10):797, 2019.

[2] Jamie Snider, Max Kotlyar, Punit Saraon, Zhong Yao, Igor Jurisica, and Igor Stagljar. Fundamentals of protein interaction network mapping. *Molecular Systems Biology*, 11(12):848, 2015.

[3] Christopher A Ross and Michelle A Poirier. Protein aggregation and neurodegenerative disease. *Nature Medicine*, 10(Suppl 7):S10–S17, 2004.

[4] Valur Emilsson, Marjan Ilkov, John R. Lamb, Nancy Finkel, Elias F. Gudmundsson, Rebecca Pitts, Heather Hoover, Valborg Gudmundsdottir, Shane R. Horman, Thor Aspelund, Le Shu, Vladimir Trifonov, Sigurdur Sigurdsson, Andrei Manolescu, Jun Zhu, Örn Olafsson, Johanna Jakobsdottir, Scott A. Lesley, Jeremy To, Jia Zhang, Tamara B. Harris, Lenore J. Launer, Bin Zhang, Gudny Eiriksdottir, Xia Yang, Anthony P. Orth, Lori L. Jennings, and Vilmundur Gudnason. Co-regulatory networks of human serum proteins link genetics to disease. *Science (New York, N.Y.)*, 361, 8 2018.

[5] Jiajun Qiu Id, Kui Chen, Chunlong Zhong, Sihao Zhu, and Xiao Ma Id. Network-based protein-protein interaction prediction method maps perturbations of cancer interactome. *PLOS Genetics*, 17:e1009869, 11 2021.

[6] Feixiong Cheng, Junfei Zhao, Yang Wang, Weiqiang Lu, Zehui Liu, Yadi Zhou, William R. Martin, Ruisheng Wang, Jin Huang, Tong Hao, Hong Yue, Jing Ma, Yuan Hou, Jessica A. Castrillon, Jiansong Fang, Justin D. Lathia, Ruth A. Keri, Felice C. Lightstone, Elliott Marshall Antman, Raul Rabadan, David E. Hill, Charis Eng, Marc Vidal, and Joseph Loscalzo. Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nature Genetics 2021 53:3*, 53:342–353, 2 2021.

[7] Claudio Napoli, Giuditta Benincasa, Francesco Donatelli, and Giuseppe Ambrosio. Pre-

cision medicine in distinct heart failure phenotypes: Focus on clinical epigenetics. *American Heart Journal*, 224:113–128, 6 2020.

[8] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4, 8 2005.

[9] Daniela M Witten, Jerome H Friedman, and Noah Simon. New insights and faster computations for the graphical lasso view. *Journal of Computational and Graphical Statistics*, 2012.

[10] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 127–135, 2000.

[11] Fei Liu, Shao-Wu Zhang, Wei-Feng Guo, Ze-Gang Wei, and Luonan Chen. Inference of gene regulatory network based on local bayesian networks. *PLoS computational biology*, 12(8):e1005024, 2016.

[12] Chen Chen, Min Ren, Min Zhang, and Dabao Zhang. A two-stage penalized least squares method for constructing large systems of structural equations. *Journal of Machine Learning Research*, 19:1–34, 2018.

[13] Chunlin Li, Xiaotong Shen, and Wei Pan. Inference for a large directed acyclic graph with unspecified interventions. *arXiv preprint arXiv:2110.03805*, 2021.

[14] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. The gene ontology consortium gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[15] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[16] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[17] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[18] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[19] Timothy Shin Heng Mak et al. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41:469–480, 2017.

[20] Junghi Kim, Yun Bai, and Wei Pan. An adaptive association test for multiple phenotypes with GWAS summary statistics. *Genetic Epidemiology*, 39(8):651–663, 2015.

[21] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, 2015.

[22] Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107:223, 2012.

[23] Li Chunlin. glmtlp. `https://github.com/chunlinli/glmtlp`, 2022.

[24] Jack Pattee and Wei Pan. Penalized regression and model selection methods for polygenic scores on summary statistics. *PLOS Computational Biology*, 16(10), 2020.

[25] Renée FAG de Bruijn and M Arfan Ikram. Cardiovascular risk factors and future risk of alzheimer's disease. *BMC Medicine*, 12:1–9, 2014.

[26] Giacomo Tini, Riccardo Scagliola, Fiammetta Monacelli, Giovanni La Malfa, Italo Porto, Claudio Brunelli, and Gian Marco Rosa. Alzheimer's disease and cardiovascular disease: a particular association. *Cardiology Research and Practice*, 2020, 2020.

[27] Lasse Folkersen, Eric Fauman, Maria Sabater-Lleal, Rona J. Strawbridge, Mattias Frånberg, Bengt Sennblad, Damiano Baldassarre, Fabrizio Veglia, Steve E. Humphries, Rainer Rauramaa, Ulf de Faire, Andries J. Smit, Philippe Giral, Sudhir Kurl, Elmo Mannarino, Stefan Enroth, Åsa Johansson, Sofia Bosdotter Enroth, Stefan Gustafsson, Lars Lind, Cecilia Lindgren, Andrew P. Morris, Vilmantas Giedraitis, Angela Silveira, Anders Franco-Cereceda, Elena Tremoli, IMPROVE study group, Ulf Gyllensten, Erik Ingelsson, Søren Brunak, Per Eriksson, Daniel Ziemek, Anders Hamsten, and Anders Mälarstig. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLOS Genetics*, 13:e1006706, 4 2017.

[28] Sarah A Signor and Sergey V Nuzhdin. The evolution of gene expression in cis and trans. *Trends in Genetics*, 34(7):532–544, 2018.

[29] Daniel Swerdlow, Karoline Kuchenbaecker, Sonia Shah, Reecha Sofat, Michael Holmes, Jon White, Jennifer Mindell, Mika Kivimaki, Eric Brunner, John Whittaker, Juan Casa, and Aroon Hingorani. Selecting instruments for mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology*, 45(5):1600–1616, 2016.

[30] Gibran Hemani, Jack Bowden, and George Davey Smith. Evaluating the potential role of pleiotropy in mendelian randomization studies. *Human Molecular Genetics*, 27(2):195–208, 2018.

[31] Céline Bellenguez, Fahri Küçükali, Iris E Jansen, Luca Kleineidam, Sonia Moreno-Grau, Najaf Amin, Adam C Naj, Rafael Campos-Martin, Benjamin Grenier-Boley, Victor Andrade, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics*, 54(4):412–436, 2022.

[32] Elina M Sutinen, Minna A Korolainen, Jukka Häyrinen, Irina Alafuzoff, Steven Petratos, Antero Salminen, Hilkka Soininen, Tuula Pirttilä, and Johanna O Ojala. Interleukin-18 alters protein expressions of neurodegenerative diseases-linked proteins in human SH-SY5Y neuron-like cells. *Frontiers in Cellular Neuroscience*, 8:214, 2014.

[33] Marco Calabrò, Carmela Rinaldi, Giuseppe Santoro, and Concetta Crisafulli. The biological pathways of Alzheimer disease: A review. *AIMS Neuroscience*, 8(1):86, 2021.

[34] David et al Ochoa. Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Research*, 49(D1):D1302–D1310, 2021.

[35] Mario Barbagallo and Ligia J Dominguez. Type 2 diabetes mellitus and Alzheimer's disease. *World J Diabetes*, 5(6):889–893, 2014.

[36] Ting Li, Zheng Ning, and Xia Shen. Improved estimation of phenotypic correlations using summary association statistics. *Frontiers in Genetics*, 12:665252, 2021.

[37] Yangqing Deng and Wei Pan. Improved use of small reference panels for conditional and joint analysis with gwas summary statistics. *Genetics*, 209(2):401–408, 2018.

[38] GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[39] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[40] Gibran Hemani, Jie Zheng, Benjamin Elsworth, Kaitlin H Wade, Valeriia Haberland, Denis Baird, Charles Laurin, Stephen Burgess, Jack Bowden, Ryan Langdon, et al. The MR-Base platform supports systematic causal inference across the human phenome. *elife*, 7:e34408, 2018.