# A parametric bootstrap approach for computing confidence intervals for genetic correlations with application to genetically-determined protein-protein networks.

Yi-Ting Tsai[1,2,†], Yana Hrytsenko[1,3,4,†], Michael Elgart[1,3], Usman Tahir[3,4], Zsu-Zsu Chen[3,5], James G Wilson[3,4], Robert Gerszten[3,4], Tamar Sofer[1,2,3,4,*]

[1]Department of Medicine, Brigham and Women's Hospital, Boston, MA

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

[3]Department of Medicine, Harvard Medical School, Boston, MA

[4]CardioVascular Institute (CVI), Beth Israel Deaconess Medical Center, Boston, MA

[5]Department of Internal Medicine, Division of Endocrinology, Diabetes, and Metabolism, Beth Israel Deaconess Medical Center, Boston, MA

[†] These authors equally contributed to the work.

*Correspondence:
Tamar Sofer
Center for Life Sciences CLS-934
3 Blackfan St
Boston MA 02115
tsofer@bidmc.harvard.edu

## Abstract

Genetic correlation refers to the correlation between genetic determinants of a pair of traits. When using individual-level data, it is typically estimated based on a bivariate model specification where the correlation between the two variables is identifiable and can be estimated from a covariance model that incorporates the genetic relationship between individuals, e.g., using a pre-specified kinship matrix. Inference relying on asymptotic normality of the genetic correlation parameter estimates may be inaccurate when the sample size is low, when the genetic

correlation is close to the boundary of the parameter space, and when the heritability of at least one of the traits is low. We address this problem by developing a parametric bootstrap procedure to construct confidence intervals for genetic correlation estimates. The procedure simulates paired traits under a range of heritability and genetic correlation parameters, and it uses the population structure encapsulated by the kinship matrix. Heritabilities and genetic correlations are estimated using the close-form, method of moment, Haseman-Elston regression estimators. The proposed parametric bootstrap procedure is especially useful when genetic correlations are computed on pairs of thousands of traits measured on the same exact set of individuals. We demonstrate the parametric bootstrap approach on a proteomics dataset from the Jackson Heart Study.

## Key words

Genetic correlation; Heritability; Parametric bootstrap; Sampling; Protein-protein network.

## Introduction

Genetic correlation measures the relationship between a pair of traits through their shared genetic variability (1). It is a related concept to heritability, which measures the overall genetic contribution to a trait (2). Specifically, genetic correlation is defined as the correlation between the genetic effects of two traits. It can be estimated using individual-level data, or using summary statistics from genome-wide association studies (GWAS) (3). Scientific papers studying genetic architecture of health and behavioral phenotypes now routinely report genetic correlation estimates between phenotypes, sometimes as a step preceding follow up analysis, e.g. with polygenic risk scores or Mendelian randomization analyses (4–7). Genetic correlations are

further being studied at the local genomic region level (local genetic correlations), or stratified by genetic annotations, to localize sources of shared genetic underpinning of phenotypes (8–12).

Methods for estimating heritability and genetic correlations based on summary statistics from GWAS (3,13,14) became popular in recent years due to their computational tractability and the access to many phenotypes that were interrogated in GWAS by the research community. However, in diverse populations and in small datasets it is still preferable to estimate heritabilities and genetic correlations using individual-level data, rather than based on GWAS summary statistics (15). Methods using individual-level data typically rely on an underlying linear mixed model (LMM) formulation, where a genetic relationship matrix is used to model the relationship, or degree of similarity, between the phenotype levels of different individuals (16,17). When estimating genetic correlation between two phenotypes, a bivariate normal model is usually used. Common algorithms for estimating heritability and genetic correlation include Restricted Maximum Likelihood (REML)-based normal likelihood models (18), and method of moment estimators such as the Haseman-Elston approach (15). Estimation of standard errors (SEs), confidence intervals (CIs), and p-values, often relies on asymptotic normal approximation. However, both heritability and genetic correlations have a limited support: heritability is bounded on the [0,1] interval, and genetic correlation on the [-1,1] interval. This means that asymptotic normal approximation may not be appropriate when estimates are close to the boundary of the parameter space, and the problem is more severe with smaller datasets. Previous publications addressed the problem of confidence interval estimation in the context of heritability (19,20), but, although the distribution of genetic correlations has been studied (21–23), methods for confidence interval computation in the era of large-scale genomic studies have not been as

developed. Notably, we previously proposed a Fisher's transformation-based approach and a blocked bootstrap, relying on resampling from the data, by blocks of related individuals (15). The blocked bootstrap worked better than the Fisher's transformation approach, but was computationally more intensive and we therefore only allowed for a small number of resamples, limiting the potential coverage of the confidence intervals as well as application at scale (i.e., for millions of traits). Here, we build on a prior work by Schweiger et al. (19), in the context of heritability. We expand their parametric bootstrap test-inversion method which eliminates the dependency on asymptotic approximation.

In this paper, we develop a parametric bootstrap approach to construct CIs for genetic correlations to better model the unknown distribution of genetic correlations. The procedure requires simulating pairs of phenotypes using existing correlation structure between individuals in a given dataset, based on sets of values of heritabilities and genetic correlation between the phenotypes. The results from the simulation study are used to construct CIs for the genetic correlation parameter based on triplets of estimated heritabilities and genetic correlation of a pair of phenotypes, using the conditional empirical probability mass function (PMF) of the genetic correlation parameter. We demonstrate and compare, through simulations, the performance of two variations of the parametric bootstrap procedure, and further compare them with construction of CIs based on the Fisher's transformation of the estimated genetic correlation, and estimated standard errors (SEs) of the correlation parameter from asymptotic normal assumption on restricted maximum likelihood estimates. Despite being a resampling method, typically requiring many computations and thus computationally costly, our approach is very useful when estimating genetic correlations between thousands of traits measured on the same dataset,

because the simulation study used to construct PMFs is performed once and may be used many times. Thus, we demonstrate the application of the parametric bootstrap approach to study genetic correlations between a high-dimensional set of proteins and to develop protein-protein networks based on the genetic correlations estimated in the Jackson Heart Study dataset.

# Methods

## Linear Mixed Model (LMM) formulation

Let $\boldsymbol{y}$ be an $n \times 1$ phenotype outcome vector and $\boldsymbol{X}$ be an $n \times p$ matrix containing values of $p$ covariates measured on $n$ participants. Let $\boldsymbol{e}$ be an $n \times 1$ vector of residuals, or errors, which we assume are potentially correlated across participants due to shared genetic effects. Suppose that the $n \times n$ matrix $\boldsymbol{K}$ models the genetic relationship between individuals, such that its $i, j$ entry $k_{i,j}$ is, for example, (twice) the kinship coefficient between individual $i$ and $j$, and is equal to $k_{j,i} = k_{i,j}$ (i.e., this is a symmetric matrix). Note that genetic relationship could be estimated by various quantities (24), without loss of generality, we here assume that we use a kinship matrix using identity by descent estimates. Following standard linear mixed model formulation of heritability, we model the outcome as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e},$$

where $\boldsymbol{\beta}$ are the regression coefficients of the covariates, here treated as nuisance parameters. Suppose that the total variance is decomposed to a genetic variance and remaining residual variance. Let $\sigma_k^2$ be the genetic variance component, and $\sigma_e^2$ be the residual variance component, so that

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma_k^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I_n}).$$

The narrow-sense heritability, defined as the proportion of total variance explained by additive genetic factors is:

$$h^2 = \frac{\sigma_k^2}{\sigma_k^2 + \sigma_e^2}.$$

Here, we assume, without loss of generality, that $\sigma_k^2 + \sigma_e^2 = 1$. Therefore, we have $\sigma_k^2 = h^2$, meaning that the genetic variance is equal to the heritability. Under this assumption, the variance of the phenotype can be written as

$$var(\boldsymbol{y}) = \sigma_k^2 \boldsymbol{K} + (1 - \sigma_k^2)\boldsymbol{I_n}.$$

Given two $n \times 1$ vectors $\boldsymbol{y_1}, \boldsymbol{y_2}$, their covariance can be modelled as

$$cov(\boldsymbol{y_1}, \boldsymbol{y_2}) = \sigma_{k,1}\sigma_{k,2}\, \rho_k\, \boldsymbol{K} + \sigma_{e,1}\sigma_{e,2}\rho_e \boldsymbol{I_n},$$

where $\sigma_{k,i}^2$ is the genetic variance for phenotype $i \in \{1,2\}$, $\sigma_{e,i}^2$ is the residual variance for phenotype $i$, $\rho_k$ is the genetic correlation between the two phenotypes, and $\rho_e$ is the residual correlation between the two phenotypes (15). Alternatively:

$$cov(\boldsymbol{y_1}, \boldsymbol{y_2}) = \sigma_{k,1}\sigma_{k,2}\, \rho_k\, \boldsymbol{K} + \sqrt{1 - \sigma_{k,1}^2}\sqrt{1 - \sigma_{k,2}^2}\, \rho_e\, \boldsymbol{I_n} \tag{1}$$

If we further plug in $\sigma_k^2 = h^2$, then, for a single and for two phenotypes, we can write the variance model as:

$$var(\boldsymbol{y}) = h^2 \boldsymbol{K} + (1 - h^2)\boldsymbol{I_n} \tag{2}$$

$$cov(\boldsymbol{y_1}, \boldsymbol{y_2}) = h_1 h_2\, \rho_k\, \boldsymbol{K} + \sqrt{1 - h_1^2}\sqrt{1 - h_2^2}\, \rho_e\, \boldsymbol{I_n} \tag{3}$$

which is the form that we will use to simulate outcomes in the following parametric bootstrap section.

## Parametric Bootstrap

We use a parametric bootstrap approach to compute confidence intervals. In brief, we simulate data for each set of potential values of heritabilities $\tilde{h}^2{}_1, \tilde{h}^2{}_2$ and genetic and residual correlation $\tilde{\rho}_k, \tilde{\rho}_e$ between two phenotypes based on the existing genetic relationship between individuals in the dataset of interest. Next, we compute confidence intervals by inferring ranges of potential values of $\rho_k$ (integrated over potential values of $h_1^2, h_2^2$, as the true values are not known) that resulted in realized (estimated) values $\hat{h}^2{}_1, \hat{h}^2{}_2, \hat{\rho}_k$. In practice, to limit computational burden, we fix $\tilde{\rho}_e = 0$ (and we assess the use of $\tilde{\rho}_e = 0.2, 0.4$ also as fixed values in simulations).

**Step 1: Random sampling of genetically correlated outcomes**

For every given combination of the potential heritability of phenotype 1 ($\tilde{h}^2{}_1$), potential heritability of phenotype 2 ($\tilde{h}^2{}_2$), and potential genetic correlation ($\tilde{\rho}_k$), we draw $N$ (e.g., 10,000) pairs of phenotype vectors $(\boldsymbol{y_1}, \boldsymbol{y_2})$ from the multivariate normal distribution

$$\begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{pmatrix} \sim N\left[\mathbf{0}_{2n\times 1}, \begin{pmatrix} var(\boldsymbol{y_1}) & cov(\boldsymbol{y_1}, \boldsymbol{y_2}) \\ cov(\boldsymbol{y_1}, \boldsymbol{y_2}) & var(\boldsymbol{y_2}) \end{pmatrix}\right],$$

where

$$\begin{cases} var(\boldsymbol{y_1}) = \tilde{h}_1^2 \boldsymbol{K} + (1-\tilde{h}_1^2)\boldsymbol{I}_n \\ var(\boldsymbol{y_2}) = \tilde{h}_2^2 \boldsymbol{K} + (1-\tilde{h}_2^2)\boldsymbol{I}_n \\ cov(\boldsymbol{y_1}, \boldsymbol{y_2}) = \tilde{h}_1\tilde{h}_2\, \tilde{\rho}_k\, \boldsymbol{K} + \sqrt{1-\tilde{h}_1^2}\sqrt{1-\tilde{h}_2^2}\ \tilde{\rho}_e\, \boldsymbol{I}_n \end{cases}$$

where $\tilde{h}^2{}_1, \tilde{h}^2{}_2 \in [0,1]$ and $\tilde{\rho}_k \in [-1,1]$. We note here that $\rho_e$ may take potential value in the interval $[-1,1]$, but we choose just one value as mentioned earlier. We used 10 settings for $\tilde{h}^2{}_1$, 10 settings for $\tilde{h}^2{}_2$, and 20 settings for $\tilde{\rho}_k$ as follows:

$$\tilde{h}_1 = \{0.05, 0.15, 0.25, \ldots, 0.85, 0.95\}$$

$$\tilde{h}_2 = \{0.05, 0.15, 0.25, \ldots, 0.85, 0.95\}$$

$$\tilde{\rho}_k = \{-0.95, -0.85, \ldots, -0.05, 0.05, 0.15, \ldots, 0.85, 0.95\}.$$

Under this setup, there are 2,000 distinct combinations of triplets $(\tilde{h}^2{}_1, \tilde{h}^2{}_2, \tilde{\rho}_k)$ in total. We note that while developing this procedure we compared using finer grids of values, with sequences with differences of 0.01 between each two consecutive values, but the results remained essentially the same while the computational burden was substantially higher. Because the grid size determines the accuracy level of potential confidence interval coverage, we later offer a solution to increase coverage without simulating a finer grid of values.

### Step 2: Genetic Correlation and Heritability Estimation

Next, based on each sampled pair of phenotype vectors $(\boldsymbol{y_1}, \boldsymbol{y_2})$ we estimate $\hat{h}_1^2, \hat{h}^2{}_2, \hat{\rho}_k$. While the procedure is in principle naïve to the specific formula used, we are using the closed-form Hasemen-Elson formulas we previously derived (15,20):

$$\hat{h}_l^2 = \frac{\boldsymbol{y}_l^T \boldsymbol{W} \boldsymbol{y}_l}{tr(\boldsymbol{W}\boldsymbol{W})}, \quad l \in \{1,2\} \tag{4}$$

$$\hat{\rho}_k = \frac{\boldsymbol{y}_1^T \boldsymbol{W} \boldsymbol{y}_2}{\sqrt{\boldsymbol{y}_1^T \boldsymbol{W} \boldsymbol{y}_1}\sqrt{\boldsymbol{y}_2^T \boldsymbol{W} \boldsymbol{y}_2}} \tag{5}$$

Where $\boldsymbol{W}$ is either the kinship matrix with all diagonal values set to zero, or, a weighted sum of the kinship matrix $\boldsymbol{K}$ and the matrix modelling the random error (here, an identity matrix) with weights related to the relationship between the kinship matrix and the identity matrix. See (15) for more details, including the potential use of multiple matrices modelling correlations between individuals. In practice, it is appropriate to use the kinship matrix with diagonal value set to zero when only the kinship matrix is used to model relationship between individuals. Using these formulas rather than likelihood-based procedures is computationally quicker as no iterations are required.

**Step 3: PMF estimation for $\Pr(\boldsymbol{\rho}_k | \widehat{\boldsymbol{h}}_1^2, \widehat{\boldsymbol{h}}_2^2, \widehat{\boldsymbol{\rho}}_k^2)$**

We now derive the expression for the conditional probability of $\rho_k$ given the estimated parameters. Because the support of $h_1^2, h_2^2, \rho_k$ are continuous where $h_1, h_2 \in [0, 1]$ and $\rho_k \in [-1, 1]$, while the results from simulations are discrete values, we divide these ranges into bins, e.g., of size 0.1, i.e., forcing them into a discrete distribution:

$\rho_k \in \{\mathcal{A}_1^\rho = [-1, -0.9), \mathcal{A}_2^\rho = [-0.9, -0.8), \ldots, \mathcal{A}_{19}^\rho = [0.8, 0.9), \mathcal{A}_{20}^\rho = [0.9, 1]\}$.

$h_1, h_2 \in \{\mathcal{A}_1^h = [0, 0.1), \mathcal{A}_2^h = [0.1, 0.2), \ldots, \mathcal{A}_9^h = [0.8, 0.9), \mathcal{A}_{10}^h = [0.9, 1]\}$.

When estimating CIs for genetic correlations, we are given the estimates $\hat{h}_1^2, \hat{h}_2^2, \hat{\rho}_k$ and we want to identify a region $\mathcal{A}$ such that $\Pr\left(\tilde{\rho}_k \in \mathcal{A} | \hat{h}_1 \in \mathcal{A}_i^h, \hat{h}_2 \in \mathcal{A}_j^h,\ \hat{\rho}_k \in \mathcal{A}_k^\rho\right) = 1 - \alpha$ (e.g. $1 - \alpha = 95\%$). Therefore, we want to estimate the probabilities $\Pr\ (\tilde{\rho}_k \in \mathcal{A}_i^\rho | \hat{h}^2{}_1 \in \mathcal{A}_j^h, \hat{h}^2{}_2 \in \mathcal{A}_k^h, \hat{\rho}_k \in \mathcal{A}_l^\rho)$ for $i = 1, \ldots, 20$ in order to create an empirical probability mass function (PMF) and use it to generate confidence intervals, which can be derived using Bayes theorem. The derivation below uses the probabilities $\Pr\left(\hat{h}^2{}_1 \in \mathcal{A}_i^h \middle| \tilde{h}^2{}_1 \in \mathcal{A}_j^h\right), \Pr\left(\hat{h}^2{}_2 \in \mathcal{A}_i^h \middle| \tilde{h}^2{}_2 \in \mathcal{A}_j^h\right), \Pr\ (\hat{\rho}_k \in \mathcal{A}_i^\rho | \tilde{h}^2{}_1 \in \mathcal{A}_j^h, \tilde{h}^2{}_2 \in \mathcal{A}_k^h, \tilde{\rho}_k \in \mathcal{A}_l^\rho)$, which are the probabilities of $\hat{h}_1, \hat{h}_2, \hat{\rho}_k$ being in given regions conditional on the fixed values of $\tilde{h}^2{}_1, \tilde{h}^2{}_2, \tilde{\rho}_k$ (note that the probabilities of the estimated heritabilities do not depend heritabilities of other traits on of the genetic correlation between them). Moving forward, we drop the notations showing that values refer to bins (regions) for brevity, with the understanding that all probabilities refer to parameters being in bins. Therefore, we will denote $\Pr\ (\tilde{\rho}_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)$ instead of $\Pr\ (\tilde{\rho}_k \in \mathcal{A}_i^\rho | \hat{\rho}_k \in \mathcal{A}_j^\rho, \hat{h}^2{}_1 \in \mathcal{A}_k^h, \hat{h}^2{}_2 \in \mathcal{A}_l^h)$, etc.

We first note that $\Pr(\tilde{\rho}_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2) = \frac{\Pr(\tilde{\rho}_k, \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)}{\Pr(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)}$, and therefore, we need to estimate $\Pr(\tilde{\rho}_k, \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)$ and $\Pr(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)$.

**Estimating $\Pr(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)$**

We estimate $\Pr(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)$ based on the following:

$$\Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2\right) = \sum_{\tilde{\rho}_k} \sum_{\tilde{h}_1} \sum_{\tilde{h}_2} \Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2, \tilde{\rho}_k, \tilde{h}^2{}_1, \tilde{h}^2{}_2\right) =$$

$$= \sum_{\tilde{\rho}_k} \sum_{\tilde{h}_1} \sum_{\tilde{h}_2} \Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2 \middle| \tilde{\rho}_k, \tilde{h}^2{}_1, \tilde{h}^2{}_2\right) \Pr\ (\tilde{\rho}_k, \tilde{h}^2{}_1, \tilde{h}^2{}_2) =$$

$$= \frac{1}{n_\rho n_{\tilde{h}}^2} \sum_{\tilde{\rho}_k} \sum_{\tilde{h}_1} \sum_{\tilde{h}_2} \Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2 \mid \tilde{\rho}_k, \tilde{h}^2{}_1, \tilde{h}^2{}_2\right),$$

where, for bins of length 0.1, $n_\rho = 20$, $n_h = 10$.

**Estimating $\Pr\left(\tilde{\rho}_k, \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2\right)$**

We estimate $\Pr(\tilde{\rho}_k, \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)$ based on the following:

$$\Pr\left(\tilde{\rho}_k, \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2\right) = \Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2 | \tilde{\rho}_k\right) \Pr\left(\tilde{\rho}_k\right) = \frac{1}{n_\rho} \Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2 | \tilde{\rho}_k\right) =$$

$$= \frac{1}{n_\rho} \sum_{\tilde{h}_1} \sum_{\tilde{h}_2} \Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2, \tilde{h}^2{}_1, \tilde{h}^2{}_2 \middle| \tilde{\rho}_k\right) =$$

$$= \frac{1}{n_\rho} \sum_{\tilde{h}_1} \sum_{\tilde{h}_2} \Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2 \middle| \tilde{\rho}_k, \tilde{h}^2{}_1, \tilde{h}^2{}_2\right) \Pr\left(\tilde{h}^2{}_1, \tilde{h}^2{}_2 | \tilde{\rho}_k\right) =$$

$$= \frac{1}{n_\rho} \sum_{\tilde{h}_1} \sum_{\tilde{h}_2} \Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2 \middle| \tilde{\rho}_k, \tilde{h}^2{}_1, \tilde{h}^2{}_2\right) \Pr\left(\tilde{h}^2{}_1\right) \Pr\left(\tilde{h}^2{}_2\right) =$$

$$= \frac{1}{n_\rho n_h^2} \sum_{\tilde{h}_1} \sum_{\tilde{h}_2} \Pr(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2 | \tilde{\rho}_k, \tilde{h}^2{}_1, \tilde{h}^2{}_2)$$

Putting these together:

$$\Pr(\tilde{\rho}_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2) = \frac{\Pr(\tilde{\rho}_k, \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)}{\Pr(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)} = \frac{\sum_{\tilde{h}_1} \sum_{\tilde{h}_2} \Pr(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2 | \tilde{\rho}_k, \tilde{h}^2{}_1, \tilde{h}^2{}_2)}{\sum_{\tilde{\rho}} \sum_{\tilde{h}_1} \sum_{\tilde{h}_2} \Pr\left(\hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2 | \tilde{\rho}_k, \tilde{h}^2{}_1, \tilde{h}^2{}_2\right)}$$

$\Pr(\rho_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)$ (computed for each pre-defined region) is then the empirical probability mass function of $\rho_k$ obtained by parametric bootstrap.

**Computing confidence intervals from the PMF**

After obtaining the empirical PMF from parametric bootstrap, we can now derive the CIs for any given genetic correlation estimate $\hat{\rho}_k$ with a coverage probability of $1-\alpha$ (e.g., 95%). Because the parameters are bounded, constructed confidence intervals may be asymmetric in both the distance between the estimated $\hat{\rho}_k$ to the low and high values of the confidence interval, and in the cumulative probability between provided by the two "sides" (around $\hat{\rho}_k$) of the confidence interval. We address this by considering the following three cases depending on the cumulative probability

$$cp_l = \sum_{\rho_k \leq \hat{\rho}_k} \Pr(\rho_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)\,.$$

Here $cp_l$ denotes cumulative probability of potential $\rho_k$ values lower or equal to the estimated $\hat{\rho}_k$. Denote the low and the high values of the confidence interval for $\hat{\rho}_k$ by $\hat{\rho}_L$ and $\hat{\rho}_H$. Then a 1-$\alpha$ confidence interval news to include all potential values $\tilde{\rho}_k$ such that:

$$\sum_{\hat{\rho}_L \leq \tilde{\rho}_k \leq \hat{\rho}_H} \Pr(\tilde{\rho}_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2)\,. \tag{6}$$

**Case 1:** If $cp_l < \frac{1-\alpha}{2}$

Here, $\hat{\rho}_L$ corresponds to the first potential value $\tilde{\rho}_k$ (i.e. a point in the first considered bin, where bins are considered by order $\mathcal{A}_1^\rho, \mathcal{A}_2^\rho$ ...) where $\Pr(\tilde{\rho}_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2) > 0$. $\hat{\rho}_H$ corresponds to the smallest potential value $\tilde{\rho}_k$ satisfying equation (6).

**Case 2:** If $1 - cp_l < \frac{1-\alpha}{2}$

In this case we first identify $\hat{\rho}_H$ as the highest $\tilde{\rho}_k$ (i.e. a point in first considered bin, where bins are considered by order $\mathcal{A}^{\rho}_{n_h}, \mathcal{A}^{\rho}_{n_h-1}$ ...) with $\Pr(\tilde{\rho}_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2) > 0$. $\hat{\rho}_L$ corresponds to the highest potential value $\tilde{\rho}_k$ satisfying equation (6).

**Case 3:** Both $cp_l > \frac{1-\alpha}{2}$ and $1 - cp_l > \frac{1-\alpha}{2}$

Here we require $\hat{\rho}_L$ to be the largest value and $\hat{\rho}_H$ to be the lowest such that

$$\sum_{\hat{\rho}_L \leq \tilde{\rho}_k} \Pr(\tilde{\rho}_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2), \sum_{\tilde{\rho}_k \geq \hat{\rho}_H} \Pr(\tilde{\rho}_k | \hat{\rho}_k, \hat{h}^2{}_1, \hat{h}^2{}_2) \geq \frac{1-\alpha}{2}.$$

When the upper bound or lower bound of CIs obtained from the above procedure lies somewhere inside the bins defined by the grid of considered values, which is often the case, we use linear interpolation to get a position for upper and lower bound as point within the bins.

**Empirical Beta Approximation to the PMF for CI estimation**

Because the PMF is discrete, it limits the potential coverage of constructed CIs and the potential computation of accurate p-values. Thus, we study a continuous beta approximation to the empirical PMF from parametric bootstrap. Since the range of genetic correlation is [-1, 1], and the range of beta distribution is [0, 1], we first map the [-1, 1] range of genetic correlations to [0, 1] range of beta distribution through a location-scale transformation. After finding the 100*(1 − $\alpha$)% CIs of $\hat{\rho}_k$ on the beta scale using a similar approach to that reported based on the discreate PMF, we apply the inverse location-scale transformation from [0, 1] to [-1, 1] to retrieve the CIs of genetic correlations.

## The Jackson Heart Study

The Jackson Heart Study (JHS) is a longitudinal study following 5,306 individuals of African American background from Jackson Mississippi (25,26). The study population included 2,050 related and unrelated JHS participants who had whole genome sequencing (WGS) through the Trans-Omics for Precision Medicine (TOPMed) program, proteomics data, and available body mass index (BMI). The TOPMed Data Coordinating Center used TOPMed WGS data from the TOPMed freeze 8 release and computed kinship matrix, tabulating the genetic relationship between TOPMed participants. We subsetted this matrix into JHS participants. Concentration levels of 1,317 proteins were measured using the SomaScan platform (27). The JHS study was approved by Jackson State University, Tougaloo College, and the University of Mississippi Medical Center Institutional Review Boards, and all participants provided written informed consent.

We excluded 5 proteins with more than 80% missing values. The remaining dataset had no missing protein measurements. Protein measurements were adjusted for batch effect by rank-normalizing each protein separately in each batch and then aggregating the data across batches. Next, the protein measurements were regressed over (1) only intercept, and (2) over age, sex, and BMI. The residuals from each of these regressions were extracted and were used for estimating heritabilities and genetic correlations between all protein pairs using Haseman-Elston estimators provided in equations (4) and (5), in addition to heritabilities and genetic correlations estimated using the rank-normalized protein levels (without regressing them on covariates). Also, we compared the estimates of genetic correlations to estimated Pearson correlations calculated using *stats* R package (version 3.6.2).

## Simulation Studies

We used the kinship matrix from the JHS data to perform simulations. To study methods' performance in larger sample sizes, we also created simulated datasets mimicking the JHS in which we used block matrices, with blocks being the original JHS kinship matrix using n = 2,050 individuals. We used 2 and 3 times the original sample size to form block diagonal kinship matrices with n = 4,100 and n = 6,150. We referred to simulations using the kinship matrix, and the 2- and 3-times block matrices as Setting A, B, and C. Thus, we used these kinship matrices to (1) perform simulations for the parametric bootstrap procedure, where in the primary we fix $\tilde{\rho}_e$ = 0.2 as a conservative potential high value of $\rho_e$. We also performed simulations comparing the choice of $\tilde{\rho}_e \in \{0, 0.2, 0.4\}$. Next, (2) we generate new simulated data that used the results of the parametric bootstrap simulations (1) to construct CIs. We performed 10,000 simulations for each combination of potential $(h_1^2, h_2^2, \rho_k)$, with $h_1^2, h_2^2 \in \{0.05, 0.15, \ldots, 0.95\}$ and $\rho_k \in \{-0.95, -0.85, \ldots, -0.05, 0.05, \ldots, 0.095\}$. We constructed CIs for the estimated $\hat{\rho}_k$ in each simulation.

### Comparison: four approaches of constructing CIs

We estimated the coverage and the width of the CIs constructed using a few approaches: (a) percentiles of the empirical PMF constructed using the parametric bootstrap approach; (b) beta approximation to the empirical PMF; and two existing methods: (c) Fisher's transformation; and (d) normal approximation to the distribution of the estimated genetic correlation implemented in the GCTA package (29).

The Fisher's transformation method assumes that genetic correlations follow the same distributions as Pearson correlations (30). More specifically, they are normally distributed after Fisher's transformation. For genetic correlation $\rho_k$,

$$z = \frac{1}{2}\ln\left(\frac{1+\rho_k}{1-\rho_k}\right) \sim \mathrm{N}(\mu, \sigma), \text{ where}$$

$$\mu = \frac{1}{2}\ln\left(\frac{1+\rho_k}{1-\rho_k}\right), \sigma = \frac{1}{\sqrt{N_{eff}-3}},$$

With $N_{eff}$ being the "effective sample size", previous proposed to be $trace(\boldsymbol{K}^{-}\boldsymbol{K}^{-})$, with $\boldsymbol{K}^{-}$ being the kinship matrix with diagonal values set to zero (15). We can then construct the CIs of z by the standard approach assuming asymptotic normal distributions. For example, the 95% CI of z would be $[\mu - 1.96\sigma, \mu + 1.96\sigma]$. After finding the 100*$(1-\alpha)$% CIs of $\hat{\rho}_k$ on the Fisher's transformed $(z)$ scale, we apply the inverse Fisher's transform to retrieve the CIs of the genetic correlation $\rho_k$.

To compute CIs based on existing approach that rely on a normal approximation, we estimate both the genetic correlation and its standard error using the bivariate REML procedure implemented in the GCTA package. We apply the $[\hat{\rho} - 1.96\,SE, \hat{\rho} + 1.96\,SE]$ formula to construct 95% CIs. Due to the high computational resources required by GCTA, we focus only on the four scenarios when true $\rho_k$ equals {0.05, 0.15, 0.45, 0.95} with the original-size kinship matrix.

**Performance evaluation metrics**

We used coverage probabilities and CI widths as the metrics to evaluate and compare the performance of the four approaches for CI construction. In primary results, for a given true value

of genetic correlation $\rho_k$ we calculated both the coverage probability and the average width of 95% CIs using the constructed CIs for the estimated $\rho_k$ over all the 100 true heritability scenarios (10 for $h_1^2$, and 10 for $h_2^2$). Ideally, the coverage probabilities would be at or above 95% across different $\rho_k$, and also having small CI widths. The coverage probability was estimated as the proportion of simulations in which the true $\rho_k$ was contained in its CI.

**P-value estimation**

We evaluated the use of the CI inversion method to obtain p-values for hypothesis testing. Here, our null hypothesis $H_0$ is $\rho_k = 0$, and our alternative hypothesis $H_1$ is $\rho_k \neq 0$. Given any realization $(\hat{h}_1, \hat{h}_2, \hat{\rho})$, we can estimate the CIs for $\rho_k$ using the parametric bootstrap procedure, focusing on the continuous beta approximation to the empirical PMF because smaller accurate p-values can be obtained if the underlying distribution is continuous. To determine the p-values of genetic correlation estimates, we use the CI inversion methods. Suppose that we construct a $100\times(1-\alpha)\%$ CI. Then, we can determine that the p-value is smaller than $\alpha$ if the constructed CI does not cover 0. For computational efficiency, we implemented a method that constructs CIs using a binary search approach to the $\alpha$ value, stopping when a pre-defined sensitivity level is reached.

**Type 1 error**

For each combination of potential heritability values $(\tilde{h}_1^2, \tilde{h}_2^2)$, we simulated 10,000 pairs of phenotype vectors $(\boldsymbol{y_1}, \boldsymbol{y_2})$ under the null, i.e., $\rho = 0$, estimated their genetic correlations, and calculated their p-values as described above based on the beta approximation to the PMF. After obtaining the p-values for all the 10,000 simulated data, we estimated the type 1 error rate, also

called the size of the test, by checking the percentage of these simulation rejecting the null given an $\alpha$ value.

# Results

## Simulation studies

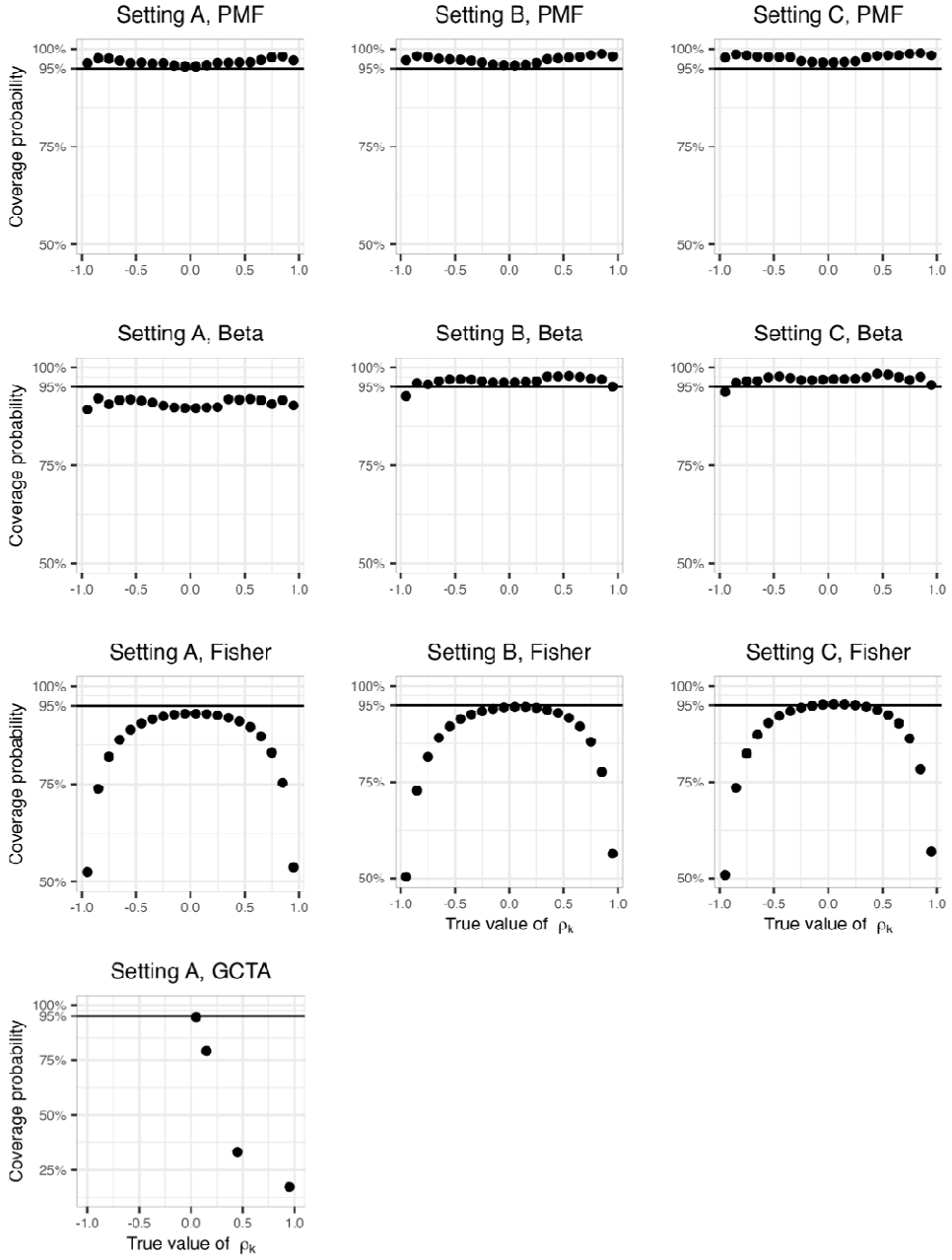
In primary results, for a given true value of genetic correlation $\rho_k$ we calculated both the coverage probability and the average width of 95% CIs by averaging the corresponding estimates over all the 100 true heritability scenarios (10 for $h_1^2$, and 10 for $h_2^2$). Supplementary Table 1 provides estimated coverage by all combinations of (true) $\rho_k, h_1^2, h_2^2$. Figure 1 provides the estimated coverage probabilities for the compared methods in simulations, and Figure 2 provides the averaged CI widths. The PMF approach provides appropriate coverage across the three settings defined by the kinship matrices. The beta approximation to the PMF results in under-coverage across the simulated $\rho_k$ values setting A, but improved substantially in settings B and C when the simulated sample size increased. Still the average width of the CIs was lower when using the empirical PMF. In setting A, GCTA had an appropriate coverage only when $\rho_k$ was set to 0.05. The Fisher's transformation tended to result in under-coverage.

Figure 3 compares coverage and CI widths when using the empirical PMF approach to compute CIs and settings $\rho_e = 0, 0.2,$ or $0.4$ in the simulations generating data. It demonstrates that there is essentially no difference in the results.

We also used the simulations to estimate type 1 error when using the confidence interval inversion methods with the beta approximation to the PMF to compute association p-values. The

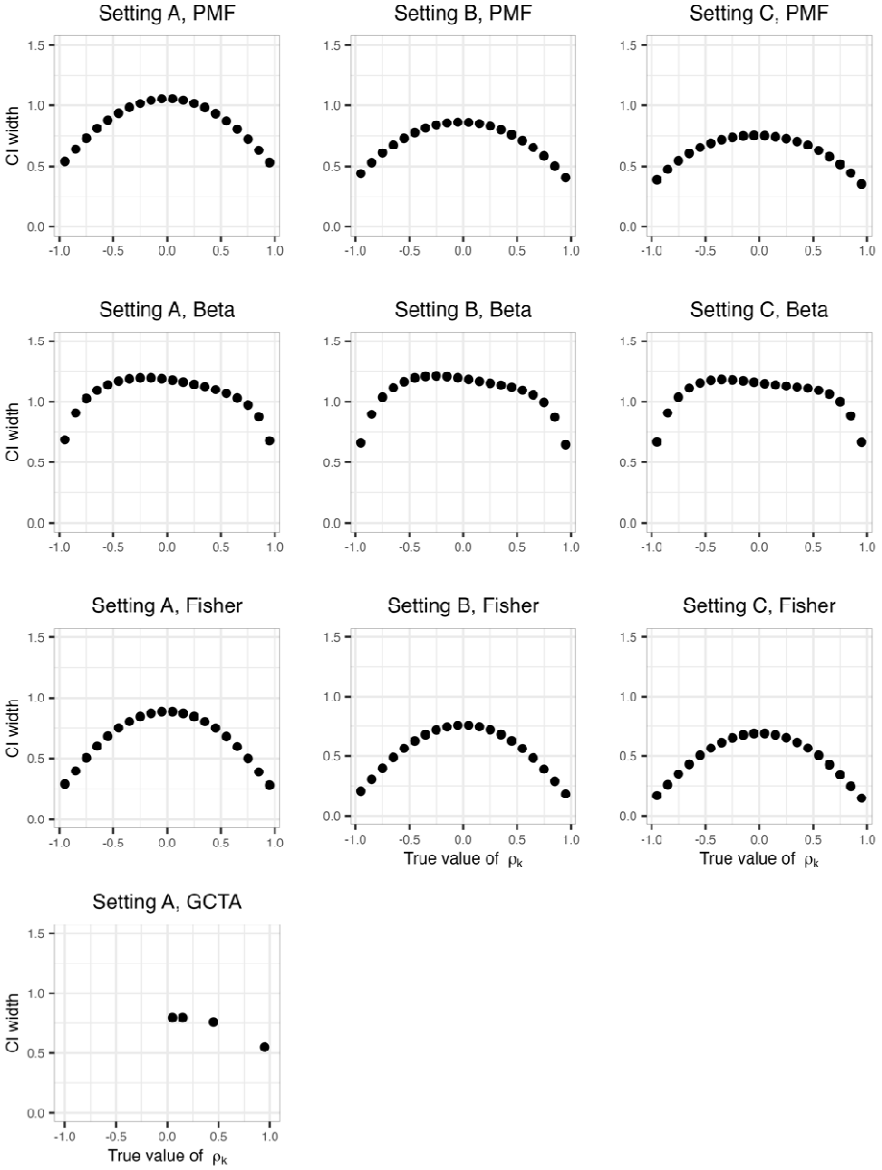
results are visualized in Figure 4. Here, we also estimated type 1 error by combinations of specific $h_1^2, h_2^2$ values. With $\alpha = 0.05$, the type 1 error was controlled across heritability combinations in settings B and C, but not in setting A. While it is unsurprising that the type 1 error is not controlled when heritability values of either one of the two traits are very small, the test was also somewhat inflated in setting A when the two heritabilities were fairly high. Over all, the beta approximation method to the PMF is promising for computing high coverage CIs and p-values when the sample size is sufficiently large.

## Figure 1: Estimated coverage of 95% confidence intervals of genetic correlations in the primary simulations.



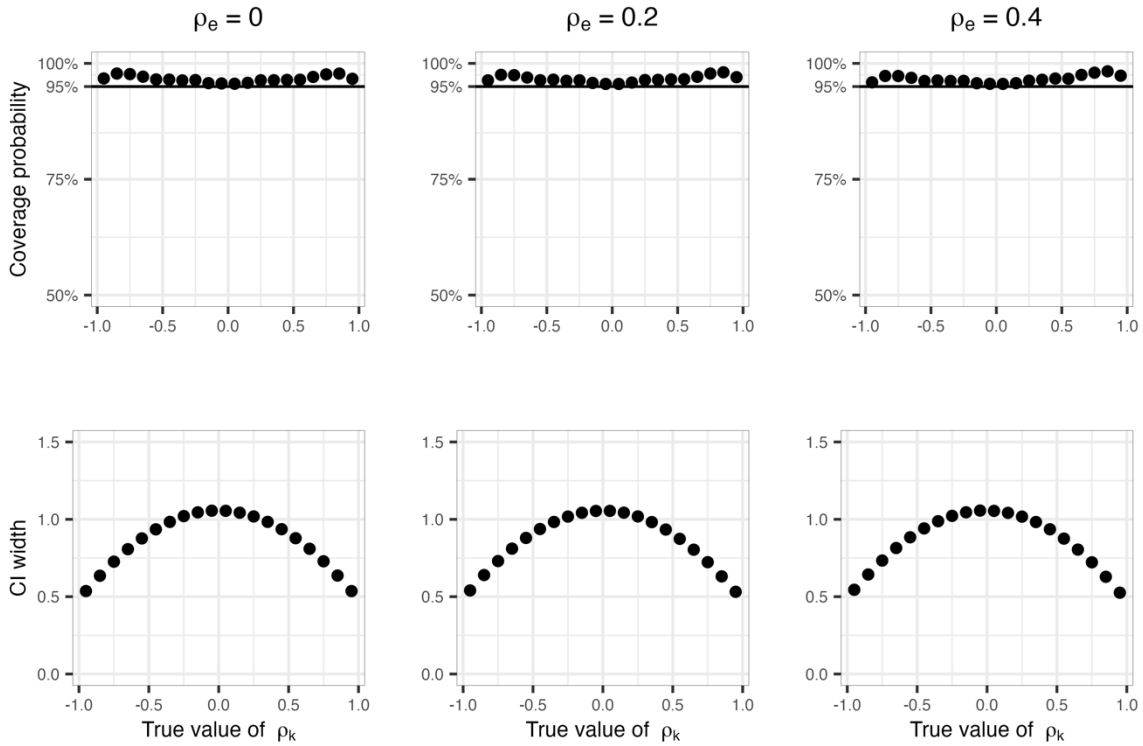

The columns represent different kinship matrix sizes: Setting A denotes the use of original-size kinship matrix (n=2,050), Setting B denotes the use of the double-size kinship matrix (n=4,100), and Setting C denotes the use of the triple-size kinship matrix (n=6,150). The rows represent the four approaches for constructing CIs, including parametric bootstrap PMF, beta approximation for parametric bootstrap PMF, Fisher's transformation, and GCTA package use of normal distribution approximation. Only parts of the analyses were carried out on the GCTA package due to the high computational resources required.

## Figure 2: Average width of the confidence intervals of the genetic correlations in the primary simulations.



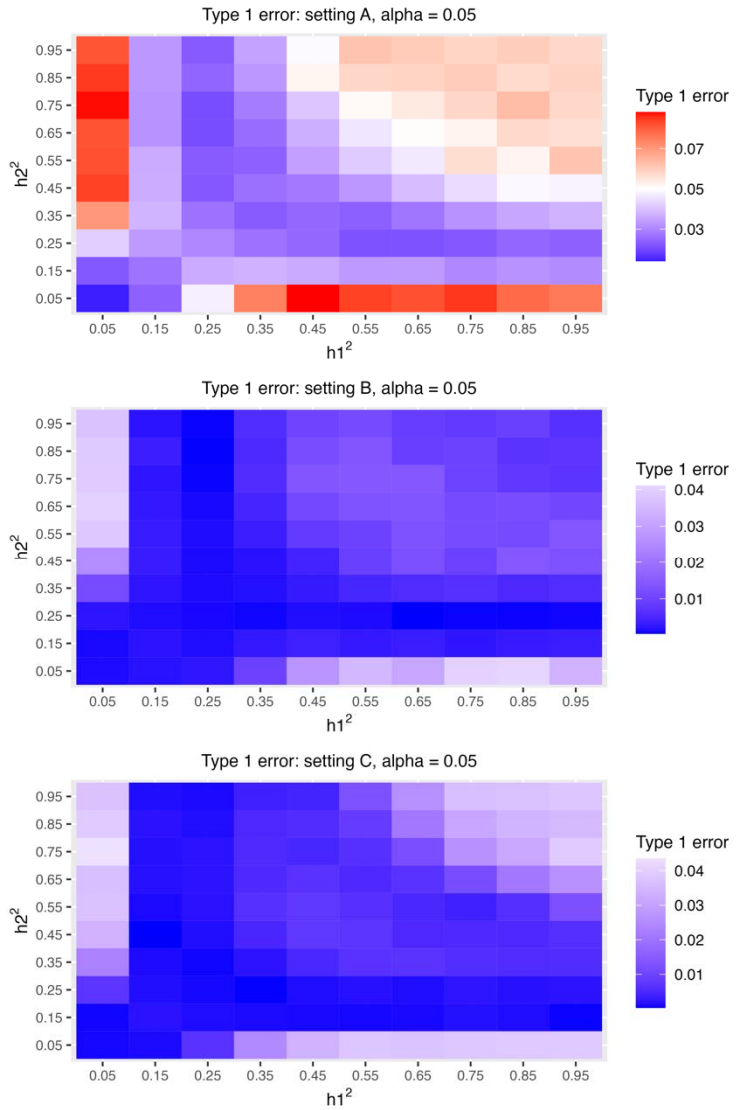

The columns represent different kinship matrix sizes: Setting A denotes the use of original-size kinship matrix (n=2,050), Setting B denotes the use of the double-size kinship matrix (n=4,100), and Setting C denotes the use of the triple-size kinship matrix (n=6,150). The rows represent the four approaches for constructing CIs, including parametric bootstrap PMF, beta approximation for parametric bootstrap PMF, Fisher's transformation, and GCTA package. Only parts of the analyses were carried out on the GCTA package due to the high computational resources required.

Figure 3: Setting environmental correlation between error terms has no effect on genetic correlation estimates.




Comparison of coverage and CI widths when using the empirical PMF approach to compute CIs and settings or in the simulations generating data.

## Figure 4: Type 1 error estimates when using the confidence interval inversion method and the Beta approximation to perform association testing




Visualization of type 1 error ( ) when using the CI inversion approach coupled with the Beta approximation to the PMF to generate CIs. The results are provided for each simulation settings and by values of .

## Application to genetically-determined protein-protein networks in JHS

We estimated heritabilities and genetic correlations for every pair of proteins among the 1,317 proteins available in JHS, in an analysis adjusted to age, sex, and BMI (in which protein measures were first regressed over these covariates prior to estimation of genetic correlations based on the resulting residuals), and in an unadjusted analysis. Characteristics of the JHS dataset are provided in Table 1. Of the study participants, 61% were women. Individuals were 55 (male)-56 (female) on average, and were mostly overweight. Some individuals were close relatives. For example, there were 341 pairs of individuals with estimated coefficient of relationship $\geq 0.48$, and 1,113 pairs of individuals with coefficient of relationship$\geq 0.12$ (considering the total number of unique pairs of individuals, this corresponds to 0.05% of all pairs of participants).

## Table 1: JHS dataset characteristics stratified by sex.

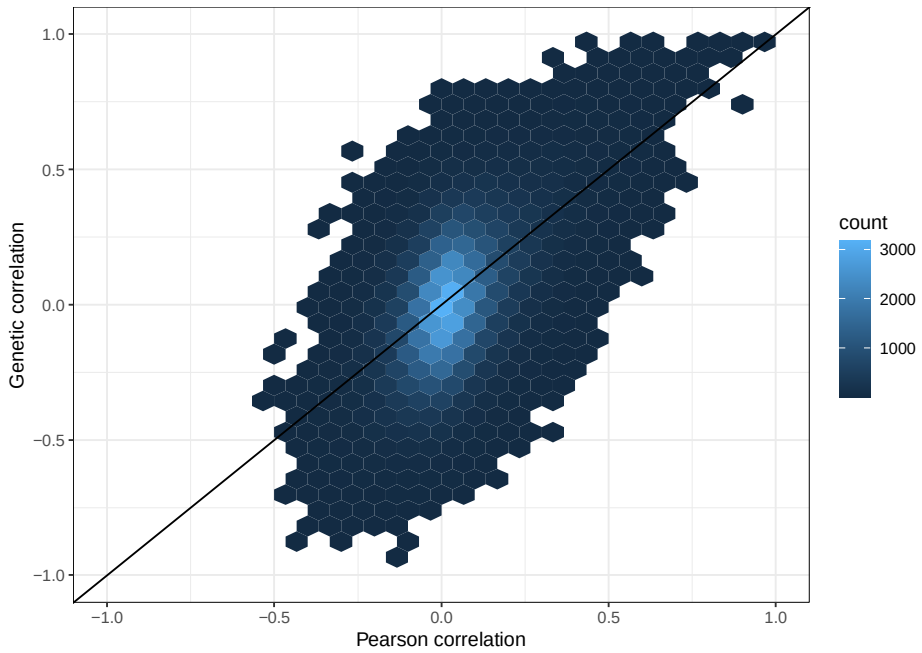
| Characteristic | Female | Male |
|---|---|---|
| **N** | 1,252 | 798 |
| **Age**[1] | 57 (46, 65) | 55 (45, 65) |
| **BMI**[1] | 32 (27, 37) | 29 (26, 32) |

[1]Median (IQR)

Supplementary Tables 2 and 3 provide the estimated heritabilities of all proteins in the dataset from analysis unadjusted and adjusted to covariates (age, sex and BMI) respectively. Based on the simulations using this specific dataset, we removed from consideration proteins with

estimated heritabilities $\hat{h}^2 < 0.3$, as genetic correlations and p-values using the beta approximation method are less reliable compared to higher values of (real, not estimated) heritabilities. We also excluded proteins with estimated $\hat{h}^2 > 0.9$, because such high may suggest a problem with the measurement and/or genetic characterization of a protein (e.g., technical issue with the platform, genetic variants segregated to a few families, etc.). After the above filtering, there were 403 and 431 proteins, or 81,406 and 93,096 protein-protein pairs, available for genetic correlation analysis in the covariate-adjusted and unadjusted analyses, respectively. For each set of the proteins (adjusted and unadjusted), it took around 2.5 hours to estimate the genetic correlations and around 12 minutes to construct the CIs, based on the previously-constructed parametric bootstrap reference results, for all the protein-protein pairs on a MacBook Pro laptop with an M1 chip. Full results from genetic correlation estimates for these sets of proteins are provided in Supplementary Tables 4 and 5. Figure 5 visualizes the comparison between estimated phenotypic (Pearson) and genetic correlations across these phenotype pairs. The figure suggests that, for this set of highly-heritable proteins, genetic correlations tend to be higher than Pearson correlations (to see this, one needs to focus in Figure 5 on the bright hexbins because they represent many more protein pairs compared to dark hexbins).

## Figure 5: Estimated Pearson versus genetic correlations between heritable proteins.




The figure compares the sample Pearson correlation to the estimated genetic correlation $\hat{\rho}_k$ for all protein pairs for which the estimated heritability $\hat{h}^2 > 0.3$ for each of the proteins. The color of each hexbin represents the number (count) of protein pairs with x- and y- axis values falling under the hexbin.

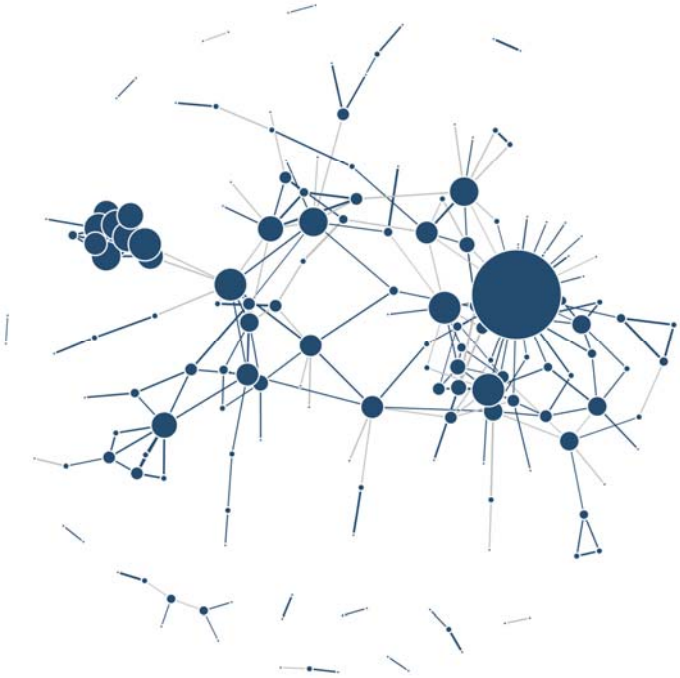
**Protein-Protein Network**

We visualize the results in a protein-protein network. Due to the large number of protein pairs, we focused the network resulting from protein-protein genetic correlations passing a p-value threshold. We computed p-values for the genetic correlations between the limited set of heritable and "valid" proteins (with heritability estimates that are not egregiously high) using the beta approximation to the PMF, and applied a False Discovery Rate (FDR) correction using the Benjamini-Hochberg procedure (31). The considered pairs of proteins are those with FDR-adjusted genetic correlation p-value<0.01. This corresponds to 253 and 294 pairs of genetically-correlated proteins in adjusted and unadjusted analysis, respectively. Figure 6 visualizes these results. The size of each node represents its degree, with larger ones being "hub nodes/proteins",

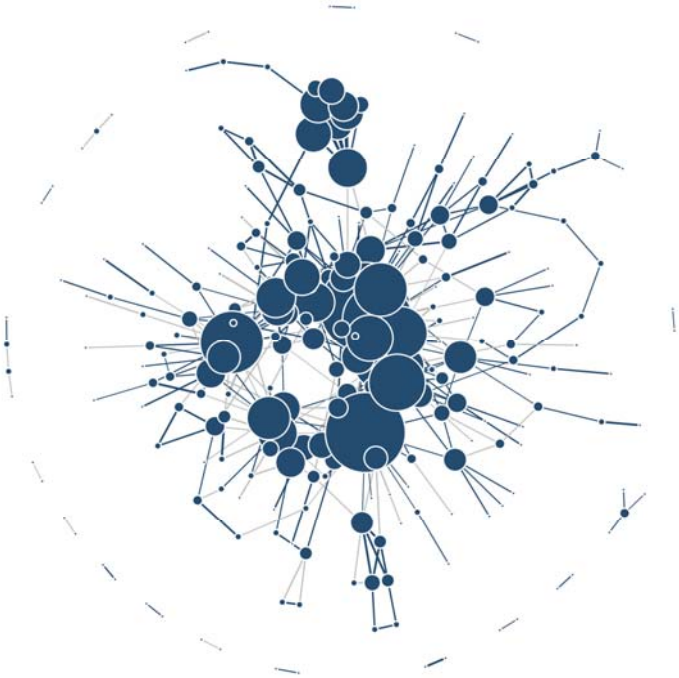
(genetically) associated with a large number of proteins. See Supplementary Tables 6 and 7 for estimated genetic correlations between pairs of proteins selected based on the criteria described above. Supplementary Table 8 contains a list of the top 10 hub nodes/proteins and their connections, i.e., the list of proteins connected to each of these hub nodes, both in covariate adjusted and unadjusted analyses. Visually, the network appears to be less connected (and we also know that the number of connections decreased) in analysis that adjusted for age, sex, and BMI. It is likely that genetic correlations decreased because BMI has strong effects on proteins, and the genetic effects on BMI are also strong, so when BMI was adjusted for, genetic effects inducing correlations between proteins were reduced.

Figure 6: Network constructed from top pairs of genetically-correlated proteins.

**a**



**b**



Panel (a) visualizes the protein-protein genetic correlation network using the age, sex, BMI-adjusted proteins; panel (b) visualizes the corresponding network based on unadjusted analysis. The blue edges represent positive genetic correlations, and the grey edges represent negative genetic correlations. Larger nodes are hub proteins where multiple proteins have strong genetic correlations with each other both in covariate adjusted and unadjusted analyses. Some of the hub proteins include *APO_D, PREKALLIKREIN, NOTCH_3, HPV_E7_TYPE18, CARBONIC_ANHYDRASE_IV, CDK5_P35, DKK_4, PAK3, TRKC, MIS, C5A , OMD, JAG1, HEPARIN_COFACTOR_II, BFGF_R, and MMP_2, GDF_11_8.*

## Discussion

We developed a parametric bootstrap procedure to estimate confidence intervals for the genetic correlation estimator, studied it in simulations, and applied it to learn a protein-protein network using a set of heritable proteins measured in the Jackson Heart Study. Our bootstrap procedure was inspired by a similar approach developed for heritability confidence intervals (19). Compared to the previous publication focusing on heritability, our approach is complicated by the need to simulate pairs of traits, including their heritabilities and genetic correlation between them, i.e., a grid of three parameters rather than one. Indeed, confidence intervals for genetic correlation depend on trait heritability, and are wider when at least one of the traits has low heritability. Thus, the computation burden of our procedure is higher. Especially, it is important to recognize that this procedure, like that of Schweiger et al., is dataset dependent, because it uses the kinship matrix of the specific dataset. However, our procedure is realistic and useful when many genetic correlations are estimated for the same dataset, as in this work. In this case the parametric bootstrap simulation step is performed once but is applied many times. A limitation for the high dimensional number of parameters (many genetic correlation parameters) is the limited level of coverage due to the discreteness of the bootstrap procedure: we cannot use the estimated conditional PMF of $\rho_k$ (conditional on the estimated genetic correlation and

heritabilities) as it is to obtain confidence intervals at the 1-$\alpha$ level when $\alpha$ is very small (e.g., $10^{-7}$). To address this, we proposed the beta approximation, after local-scale transformation, to the PMF. The beta distribution has two parameters that can be fit to many distribution functions that are on a bounded interval. Based on our simulations, CIs based on the beta approximation tend to be wider than those using the PMF directly, and they can still undercover the desired distribution in low sample sizes. However, for larger sample sizes their performance improves. Overall, we think that for larger sample sizes, e.g., 6,000 individuals, the beta approximation to the PMF will be very useful in providing reliable confidence intervals and, using the inversion method, p-values. It is important to point out that while we performed simulations with a "triple size" JHS kinship matrix, i.e., of n=6,150 individuals, the effective sample size corresponding to it is much lower than that of real potential datasets with 6,150 individuals. That is because we simulated a block diagonal matrix. Realistic kinship matrices will have non-zero off diagonal values throughout (unless forced to be zero for computational efficiency purposes (32)).

Existing methods that compute confidence intervals for genetic correlations typically utilize an asymptotic normal distribution argument, at either the untransformed or Fisher's transformation level. This is appropriate depending on the combination of four factors: sample size, underlying (true) heritabilities of each of the pair of traits, and the underlying genetic correlation. For any given pair of traits and a dataset, any one of these factors may be suboptimal, potentially leading to poor performance of confidence intervals that rely on asymptotic normality. The bootstrap procedure addresses this shortcoming. However, this procedure too does not produce perfect confidence intervals: for low values of heritability of either one of the two traits, the coverage may still be lower than desired in low sample sizes. Note that in reality we do not know the true

heritability, we only have estimated heritability. Therefore, we cannot tell whether a CI may not be reliable according to the values of the estimated heritabilities. That is why our main results are provided at an aggregate level, across simulated values of potential heritabilities.

We demonstrated the use of genetic correlations to infer genetically-determined protein-protein networks. However, we acknowledge that our analysis is limited by the relatively low sample size, which led to posing a stringent filter requiring at least 0.3 protein heritability for inclusion in the downstream analysis. While we chose to include only edges with estimated FDR-adjusted p-value<0.01 (with p-values estimated using the beta approximation), other statistical network approaches may generate sparsity using penalized multivariant regression techniques (33,34). It would be interesting to extend such approaches to genetic (rather than phenotypic) correlation-based networks. In future work we will apply the existing framework on larger datasets and develop approximation methods to further speed up the simulations required for the parametric bootstrap and the estimation of heritabilities and genetic correlations, for example, following (35).

## Acknowledgements

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases R01DK081572. The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). The

authors also wish to thank the staffs and participants of the JHS. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

## References

1. van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of polygenic disease traits: from theory to practice. Nat Rev Genet. 2019 Oct;20(10):567–81.

2. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 2008 Apr;9(4):255–66.

3. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015 Nov;47(11):1236–41.

4. O'Connor LJ, Price AL. Distinguishing genetic correlation from causation across 52 diseases and complex traits. Nat Genet. 2018 Dec;50(12):1728–34.

5. Zhang Y, Elgart M, Kurniansyah N, Spitzer BW, Wang H, Kim D, et al. Genetic determinants of cardiometabolic and pulmonary phenotypes and obstructive sleep apnoea in HCHS/SOL. EBioMedicine. 2022 Oct;84:104288.

6. Ikeda M, Tanaka S, Saito T, Ozaki N, Kamatani Y, Iwata N. Re-evaluating classical body type theories: genetic correlation between psychiatric disorders and body mass index. Psychol Med. 2018 Jul;48(10):1745–8.

7. Kappelmann N, Arloth J, Georgakis MK, Czamara D, Rost N, Ligthart S, et al. Dissecting the Association Between Inflammation, Metabolic Dysregulation, and Specific Depressive Symptoms: A Genetic Correlation and 2-Sample Mendelian Randomization Study. JAMA Psychiatry. 2021 Feb 1;78(2):161–70.

8. Shi H, Mancuso N, Spendlove S, Pasaniuc B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. Am J Hum Genet. 2017 Nov 2;101(5):737–51.

9. Zhang Y, Lu Q, Ye Y, Huang K, Liu W, Wu Y, et al. SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. Genome Biol. 2021 Sep 7;22(1):262.

10. Guo H, Li JJ, Lu Q, Hou L. Detecting local genetic correlations with scan statistics. Nat Commun. 2021 Apr 1;12(1):2033.

11. Werme J, van der Sluis S, Posthuma D, de Leeuw CA. An integrated framework for local genetic correlation analysis. Nat Genet. 2022 Mar 14;54(3):274–82.

12. Lu Q, Li B, Ou D, Erlendsdottir M, Powles RL, Jiang T, et al. A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics. Am J Hum Genet. 2017 Dec 7;101(6):939–64.

13. Weissbrod O, Flint J, Rosset S. Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics. Am J Hum Genet. 2018 Jul 5;103(1):89–99.

14. Zhang Y, Cheng Y, Jiang W, Ye Y, Lu Q, Zhao H. Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. Brief Bioinformatics. 2021 Sep 2;22(5).

15. Elgart M, Goodman MO, Isasi C, Chen H, Morrison AC, de Vries PS, et al. Correlations between complex human phenotypes vary by genetic background, gender, and environment. Cell Reports Medicine. 2022 Dec 12;

16. Furlotte NA, Eskin E. Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. Genetics. 2015 May;200(1):59–68.

17. Lee SH, van der Werf JHJ. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics. 2016 May 1;32(9):1420–2.

18. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012 Oct 1;28(19):2540–2.

19. Schweiger R, Kaufman S, Laaksonen R, Kleber ME, März W, Eskin E, et al. Fast and accurate construction of confidence intervals for heritability. Am J Hum Genet. 2016 Jun 2;98(6):1181–92.

20. Sofer T. Confidence intervals for heritability via Haseman-Elston regression. Stat Appl Genet Mol Biol. 2017 Sep 26;16(4):259–73.

21. Brown GH. An empirical study of the distribution of the sample genetic correlation coefficient. Biometrics. 1969 Mar;25(1):63.

22. Balding DJ. Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol. 2003 May;63(3):221–30.

23. Liu BH, Knapp SJ, Birkes D. Sampling distributions, biases, variances, and confidence intervals for genetic correlations. Theor Appl Genet. 1997 Jan;94(1):8–19.

24. Legarra A. Comparing estimates of genetic variance across different relationship models. Theor Popul Biol. 2016 Feb;107:26–30.

25. Wyatt SB, Diekelmann N, Henderson F, Andrew ME, Billingsley G, Felder SH, et al. A community-driven model of research participation: the Jackson Heart Study Participant Recruitment and Retention Study. Ethn Dis. 2003;13(4):438–55.

26. Taylor HA, Wilson JG, Jones DW, Sarpong DF, Srinivasan A, Garrison RJ, et al. Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. Ethn Dis. 2005;15(4 Suppl 6):S6-4.

27. Katz DH, Tahir UA, Bick AG, Pampana A, Ngo D, Benson MD, et al. Whole genome sequence analysis of the plasma proteome in black adults provides novel insights into cardiovascular disease. Circulation. 2022 Feb;145(5):357–70.

28. van Buuren S, Groothuis-Oudshoorn K. mice□: Multivariate Imputation by Chained Equations in *R*. J Stat Softw. 2011;45(3).

29. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

30. Mudholkar GS, Chaubey YP. On the distribution of Fisher's transformation of the correlation coefficient. Communications in Statistics - Simulation and Computation. 1976 Jan;5(4):163–72.

31. 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). 1995 Jan;57(1):289–300.

32. Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, Thornton TA, et al. Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics. 2019 Dec 15;35(24):5346–8.

33. Li H, Gui J. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. Biostatistics. 2006 Apr;7(2):302–17.

34. Li B, Chuns H, Zhao H. Sparse estimation of conditional graphical models with application to gene networks. J Am Stat Assoc. 2012 Jan 1;107(497):152–67.

35. Wu Y, Burch KS, Ganna A, Pajukanta P, Pasaniuc B, Sankararaman S. Fast estimation of genetic correlation for biobank-scale data. Am J Hum Genet. 2022 Jan 6;109(1):24–32.