scientific reports

Check for updates

OPEN

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

Xuewei Cao[1], Xuexia Wang[2], Shuanglin Zhang[1] & Qiuying Sha[1]✉

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

Although genome-wide association studies (GWAS) have successfully identified thousands of single nucleotide polymorphisms (SNPs) associated with a wide range of complex human traits[1,2], there is a common limitation in which GWAS focus on only a single genetic variant with a trait at a time. This limitation may limit the power to identify clinically or biologically significant genetic associations[3]. Furthermore, many genome-wide significant genetic variants are in linkage disequilibrium (LD). Different LD patterns can cause non-replicated results of the same variant in different populations[4,5]. Therefore, several powerful gene-based statistical association tests, in which the genetic information of all genetic variants in a gene is combined to obtain more informative results, have been developed, such as the Burden Test (BT)[6], the Sequence Kernel Association Test (SKAT)[7], and the Optimized SKAT (SKATO)[8].

When individual-level genotype and phenotype data are not available, the traditional gene-based association tests, BT, SKAT, and SKATO, can be extended by using GWAS summary statistics[9]. Currently, there are many GWAS summary statistics available in public resources[10]. In GWAS summary statistics, the Z-scores of genetic variants in a gene are assumed to asymptotically follow a multivariate normal distribution with a correlation matrix among all genetic variants in a gene under the null hypothesis[11], where the correlation matrix can be estimated by LD among the genetic variants in the gene[12,13]. When individual-level data are not available, LD is usually estimated using external reference panels[14,15] (i.e., 1000 Genomes Project[16]). Due to the small sample size of reference panels used to estimate LD, statistical noise (i.e., inflated type I error rates or large numbers of false positives) often exists which needs to be accounted for[17,18]. One way to reduce the statistical noise is to correct the estimated LD by a regularization procedure[19]. In the regularization procedure, a statistical white Gaussian noise is added to the LD matrix which is estimated by a reference panel. After correcting the estimated LD by the regularization procedure, we can assume that, under the null hypothesis, the Z-scores from GWAS summary

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931, USA. [2]Department of Mathematics, University of North Texas, Denton, TX, USA. ✉email: qsha@mtu.edu

| https://doi.org/10.1038/s41598-022-07465-0

statistics asymptotically follow a multivariate normal distribution with the correlation matrix being the corrected LD matrix among the genetic variants in a gene.

To increase statistical power in identifying genes that are associated with complex diseases, PrediXcan[20] and transcriptome-wide association study[12,21] (TWAS) were developed by incorporating expression quantitative trait locus (eQTL) data into GWAS. As pointed out by Zhang et al.[15], PrediXcan and TWAS can be viewed as a simple weighted linear combination of genetic variants with an eQTL—derived weight. In fact, the genetic architecture of complex traits is rarely known in advance and is likely to vary from one region to another across the genome and from one trait to another[15]. Therefore, only considering one single eQTL—derived weight, such as in PrediXcan and TWAS, may lose statistical power in identifying significant genes. Zhang et al.[15] developed an omnibus test (OT) using Cauchy combination method to integrate association evidence obtained by BT, SKAT, and SKATO based on GWAS summary data with multiple eQTL-derived weights. They showed that OT using multiple eQTL—derived weights had some potential advantages.

Inspired by the advantage of OT, in this paper, we propose a more powerful and computationally efficient method, called Overall, to aggregate the information from three types of traditional gene-based association tests (BT, SKAT, SKATO) with multiple eQTL—derived weights using GWAS summary statistics. To combine information from the three gene-based association tests, the Overall method utilizes the extended Simes procedure[5,22]. To apply the Overall method, we first need to estimate the correlation matrix among the three gene-based association tests with eQTL—derived weights under the null hypothesis. We provide an estimation method using a replication procedure[23,24]. The replication procedure only needs to be performed once to obtain the correlation matrix for each gene. Then, the p-values of Overall can be analytically computed without using permutations. To calculate the p-values of the three types of gene-based association tests (BT, SKAT, SKATO) using GWAS summary statistics with eQTL—derived weights, we use the "sumFREGAT" package in R (https://cran.r-project.org/web/packages/sumFREGAT/index.html)[9]. Once we obtain the p-values of these three tests, the p-value of our proposed method can be easily calculated using its theoretical distribution. Extensive simulation studies show that Overall can control type I error rates well and has higher power than the comparison methods across most of the simulation settings. Similar to Zhang et al.[15], we apply our method to two schizophrenia (SCZ) and two lipids trait (HDL) GWAS summary data sets. Compared with OT and other tests, the proposed method can identify more significant genes. More interestingly, some significant genes reported by GWAS catalog are only identified by our proposed method.

Statistical models and methods

Statistical models. Consider a set of M genetic variants in a gene. Let $\mathbf{Z} = (Z_1, \ldots, Z_M)^T$ be an $M \times 1$ vector of Z-scores of the M genetic variants. Note that the Z-scores is either directly provided by publicly available GWAS summary statistics or calculated from a GWAS individual-level genotype and phenotype data set. We are interested in testing the null hypothesis H_0 that none of the genetic variants in the gene is associated with a trait, whereas the alternative hypothesis is that at least one genetic variant in the gene is associated with a trait. Following Svishcheva et al.[9], Gusev et al.[12], and Yang et al.[25], we assume $\mathbf{Z} = (Z_1, \ldots, Z_M)^T \sim \text{MVN}(\mathbf{0}, \mathbf{R})$ under the null hypothesis, where $\mathbf{R}$ is the correlation matrix among $\mathbf{Z}$, which can be estimated by LD among the genetic variants in the gene[12,13]. If individual-level data are not available, LD can be estimated using external reference panels (i.e., 1000 Genomes Project[16]). However, if the sample size of a reference panel is small, LD may not be estimated correctly so that it will induce statistical noise (i.e., inflated type I error rates or large numbers of false positives)[17,18]. One way to correct the estimated LD is to use a regularization procedure by adding a statistical white Gaussian noise[9,19]. Let $\mathbf{I}_M$ be an $M \times M$ identity matrix, and the corrected correlation matrix $\mathbf{U}$ can be defined as

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

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

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

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

Suppose that there are a total of K different eQTL—derived weights from gene expression data (i.e., Genotype-Tissue Expression (GTEx) project (https://gtexportal.org/home/)), denoted as $\hat{\mathbf{W}}_k = \text{diag}\left(\hat{W}_1^k, \ldots, \hat{W}_M^k\right)$ for $k = 0, 1, \ldots, K$, where $\hat{\mathbf{W}}_0 = \text{diag}(1, \ldots, 1)$ represents a status without using any weights. In order to avoid the influence of the scale among genetic variants within each weight, we first standardize the eQTL—derived weights $\mathbf{W}_k$ as $W_m^k = \hat{W}_m^k \Big/ \sum_{m=1}^M \left|\hat{W}_m^k\right|$ for $m = 1, \ldots, M$. Based on the kth standardized weight $\mathbf{W}_k$, the weighted Z-score $\mathbf{W}_k\mathbf{Z}$ follows a multivariate normal distribution. That is,

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

We extend the three types of gene-based association tests, BT[6], SKAT[7], and SKATO[8], to incorporate the eQTL—derived weights based on GWAS summary statistics[9,26]. For the kth eQTL—derived weight, the three gene-based test statistics can be written as

$$Q_{BT}^{k} = \left(\mathbf{Z}^{T}\mathbf{W}_{k}\mathbf{1}_{M}\right)^{2},$$
$$Q_{SKAT}^{k} = (\mathbf{W}_{k}\mathbf{Z})^{T}\mathbf{W}_{k}\mathbf{Z},$$
$$Q_{SKATO}^{k} = \min_{\rho\in[0,1]}\left\{(1-\rho)Q_{SKAT}^{k} + \rho Q_{BT}^{k}\right\},$$

where $\mathbf{1}_M$ is an $M \times 1$ vector with elements of all 1s. Under the null hypothesis, Q_{BT}^{k} follows a χ^2 distribution with 1 degree of freedom; Q_{SKAT}^{k} follows a weighted sum of χ^2 distributions with 1 degree of freedom; and Q_{SKATO}^{k} follows a mixture of χ^2 distribution[8]. The p-values of these three test statistics can be easily calculated using the "sumFREGAT" package in R (https://cran.r-project.org/web/packages/sumFREGAT/index.html)[9].

Overall method. To aggregate information from these three gene-based association tests with multiple eQTL—derived weights, we develop a novel method, called Overall, which utilizes the extended Simes procedure[5,22]. Let $p_{BT}^{k}, p_{SKAT}^{k}, p_{SKATO}^{k}$ be the p-values of BT, SKAT, SKATO with kth eQTL—derived weight, $k = 0, 1, \ldots, K$, respectively, where $k = 0$ denotes a status without using any weights. Thus, there are a total of $L = 3(K+1)$ p-values from these three gene-based association tests with different weights. Let $\left(p_{(1)}, \ldots, p_{(L)}\right)$ be a sequence of the ascending p-values, where $p_{(1)} = \min_{k=0,\ldots,K}\left\{p_{BT}^{k}, p_{SKAT}^{k}, p_{SKATO}^{k}\right\}$ and $p_{(L)} = \max_{k=0,\ldots,K}\left\{p_{BT}^{k}, p_{SKAT}^{k}, p_{SKATO}^{k}\right\}$. Overall combines these L p-values using the extended Simes procedure[5,22], and the p-value of Overall is defined as

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

where m_e is the effective number of p-values among the L gene-based association tests with multiple weights, $p_{(l)}$ is the lth element of the ascending p-values, and $m_{e(l)}$ is the effective number of p-values among the top l association tests. We use a more robust measure to obtain the effective numbers m_e and $m_{e(l)}$, which was proposed by Li et al.[5]. The values of $m_{e(l)}$ and m_e can be estimated as

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

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

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

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

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

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

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

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

Study	Tissue	# of samples	References
NTR	Peripheral blood	1247	Wright et al.[28]
YFS	Whole blood	1264	Gusev et al.[12]
METSIM	Adipose	563	Gusev et al.[12]
CMC	Brain	452	Gusev et al.[12]

Table 1. Resources of the four eQTL—derived weights used in the simulation studies.

Simulation studies

Materials and comparison methods.

In our studies, we use four data sets to obtain the eQTL—derived weights downloaded from the functional summary-based imputation website (http://gusevlab.org/projects/fusion/#reference-functionaldata). The resources to obtain the four eQTL—derived weights are listed in Table 1. For each eQTL data set, we use the weights estimated by the Best Linear Unbiased Prediction (BLUP)[27].

We compare our proposed method with three existing methods, OT[15], S-PrediXcan[29], and S-TWAS[12]. These three methods are all based on GWAS summary statistics and incorporate eQTL-derived weights. Here, we briefly introduce these three methods.

OT: For a total of K different eQTL—derived weights and the three gene-based association tests (BT, SKAT, SKATO), OT aggregates information from different weights and tests by using the Cauchy combination method[30]. The test statistic of OT is defined as $Q_{OT} = \frac{1}{3(K+1)} \sum_{k=0}^{K} \left[\tan\left\{\left(0.5 - p_{BT}^{k}\right)\pi\right\} + \tan\left\{\left(0.5 - p_{SKAT}^{k}\right)\pi\right\} + \tan\left\{\left(0.5 - p_{SKATO}^{k}\right)\pi\right\}\right]$ and the corresponding p-value of the test statistic can be approximated by $p_{OT} = \frac{1}{2} - \frac{\arctan\left(Q_{OT}\right)}{\pi}$.

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

S-TWAS: For a given eQTL-derived weight, provided by a vector $\boldsymbol{w}_k = \left(W_1^k, \ldots, W_M^k\right)^T$, the test statistic of S-TWAS is defined as $Z_{S-TWAS}^{k} = \frac{\boldsymbol{w}_k^T \cdot \boldsymbol{Z}}{\sqrt{\boldsymbol{w}_k^T \cdot \mathbf{R} \cdot \boldsymbol{w}_k}}$, where $\mathbf{R}$ is the estimated LD structure among the genetic variants in a gene and the corresponding p-value can be calculated by $p_{S-TWAS}^{k} = 2\Phi\left(-\left|Z_{S-TWAS}^{k}\right|\right)$.

The number of replications needed in estimation of $\boldsymbol{\Omega}$.

To apply our proposed method, we first need to estimate the correlation matrix of p-values, $\boldsymbol{\Omega}$, under the null hypothesis for each gene. Following the estimation procedure introduced in the method section, we generate Z-scores instead of generating individual-level genotype and phenotype data. To determine the number of replications needed in the estimation of $\boldsymbol{\Omega}$, we consider 18 genes that contain different numbers of SNPs and have different LD structures. Supplementary Table S1 gives a summary of these 18 genes. We can see from Supplementary Table S1, the number of SNPs in a gene is ranging from 23 to 359 and the average per-SNP LD score in a gene is ranging from 12.72 to 170.85. We simulate a Z-score vector from a multivariate normal distribution with mean $\mathbf{0}$ and variance–covariance matrix $\mathbf{R}$, $\boldsymbol{Z} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{R})$, where $\mathbf{R}$ is the LD matrix of each gene which can be estimated using the 1000 Genomes Project (unrelated Europeans in 1000 Genomes in Phase 3; ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/). First, we use $B = 10^4$ replications to estimate $\boldsymbol{\Omega}$ under the null hypothesis, where the estimated matrix is denoted by $\hat{\Omega}$. Then, we denote $\hat{\Omega}^0$ as the correlation matrix of p-values by using B_0 replications. We vary the value of B_0 from 16 to 5000, and test the null hypothesis that the two correlation matrices, $\hat{\Omega}^0$ and $\hat{\Omega}$, are the same by using "lavaan" package (https://CRAN.R-project.org/package=lavaan)[31]. Supplementary Figure S1 shows that the p-values for the hypothesis testing in each gene are greater than 0.05 after $B_0 = 1000$ replications for all of the 18 genes. Therefore, we recommend using 1000 replications to obtain $\hat{\Omega}$ for each gene under the null hypothesis. Consequently, 1000 replications are used in the following sessions to evaluate the type I error rates and powers of Overall.

Type I error rates.

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

We generate simulated data to mimic real lipids data which we will use in "Real data analysis" section. Gene *AGTRAP* is associated with lipids trait HDL[15], There are a total of 23 genetic variants in gene *AGTRAP.* The LD block structure of these 23 genetic variants is shown in Supplementary Fig. S2. Supplementary Figure S3 shows the estimated correlation matrix $\hat{\Omega}$ for this gene. We use 10^7 replications to evaluate type I error rates of Overall

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

Table 2. Estimated type I error rates at different significance levels with 10^7 replications. The subscript denotes BT, SKAT, and SKATO using eQTL—derived weights; CMC, METSIM, NTR, and YFS indicate the resources to obtain the eQTL—derived weights. 0 indicates the methods without using eQTL—derived weights.

for gene *AGTRAP* at $5 \times 10^{-2}, 1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}$, and 1.75×10^{-6} significance levels. With 10^7 replications, a Bonferroni corrected significance level of 1.75×10^{-6} can be reached to obtain the empirical type I error rates (i.e., for 28,625 genes in the real data analysis section, the Bonferroni corrected significance level is $0.05/28625 = 1.75 \times 10^{-6}$ at 5% significance level). We further evaluate type I rates based on the other 17 genes. To save computational time, we use 2×10^5 replications to evaluate type I error rates of Overall for the 17 genes at significance levels of $1 \times 10^{-2}, 1 \times 10^{-3}$, and 1×10^{-4}. Table 2 and Supplementary Table S2 show the estimated type I error rates of Overall under various nominal significance levels for gene *AGTRAP* and the other 17 genes, respectively. From these tables, we can see that our proposed method can control type I error rates very well at different significant levels.

Power comparison. To evaluate the performance of the Overall method, we use several simulations to compare the power of Overall with the power of OT, S-PrediXcan, S-TWAS, and three types of gene-based association tests with and without eQTL—derived weights. We use BEST to represent the test with the maximum power among the three traditional gene-based association tests with and without an eQTL—derived weight, S-TWAS.B and S-PrediXcan.B to represent the maximum power of S-TWAS and S-PrediXcan with each of the eQTL—derived weights, respectively. Following the simulation settings in Nagpal et al.[32] and Zhang et al.[15], we generate individual-level genotypes, phenotypes, and different gene expression levels using the following steps:

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

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

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

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

In our simulation studies for power comparison, we consider two genes, *AGTRAP* and *C3orf22,* from the 18 genes used in the type I error evaluation and $K = 4$ and $K = 20$ eQTL—derived weights. *AGTRAP* contains 458 haplotypes for 23 genetic variants (11 common variants and 12 rare variants; MAF ranging from 0 to 0.39775); *C3orf22* contains 295 haplotypes for 42 variants (18 common variants and 24 rare variants; MAF ranging from 0 to 0.43558). Supplementary Figure S2 shows the LD block structure of the 23 genetic variants at *AGTRAP* and the

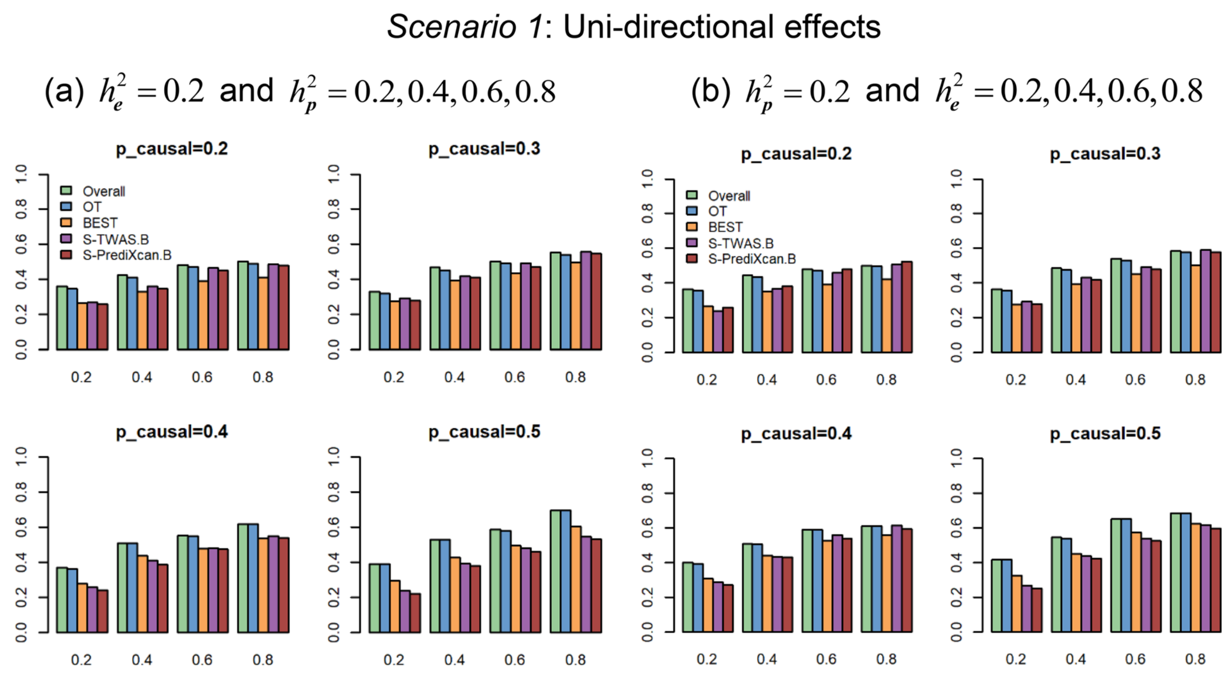

Figure 1. Power comparisons of gene-based association tests at 1.75×10^{-6} significance level under Uni-directional effects ($\beta_1 = \beta_2 = \beta_3 = \beta_4$) with $p_{causal} = (0.2, 0.3, 0.4, 0.5)$ based on gene *AGTRAP*. (**a**) Estimated power against phenotypic heritability h_p^2 with fixed expression heritability $h_e^2 = 0.2$; (**b**) Estimated power against expression heritability h_e^2 with fixed phenotypic heritability $h_p^2 = 0.2$.

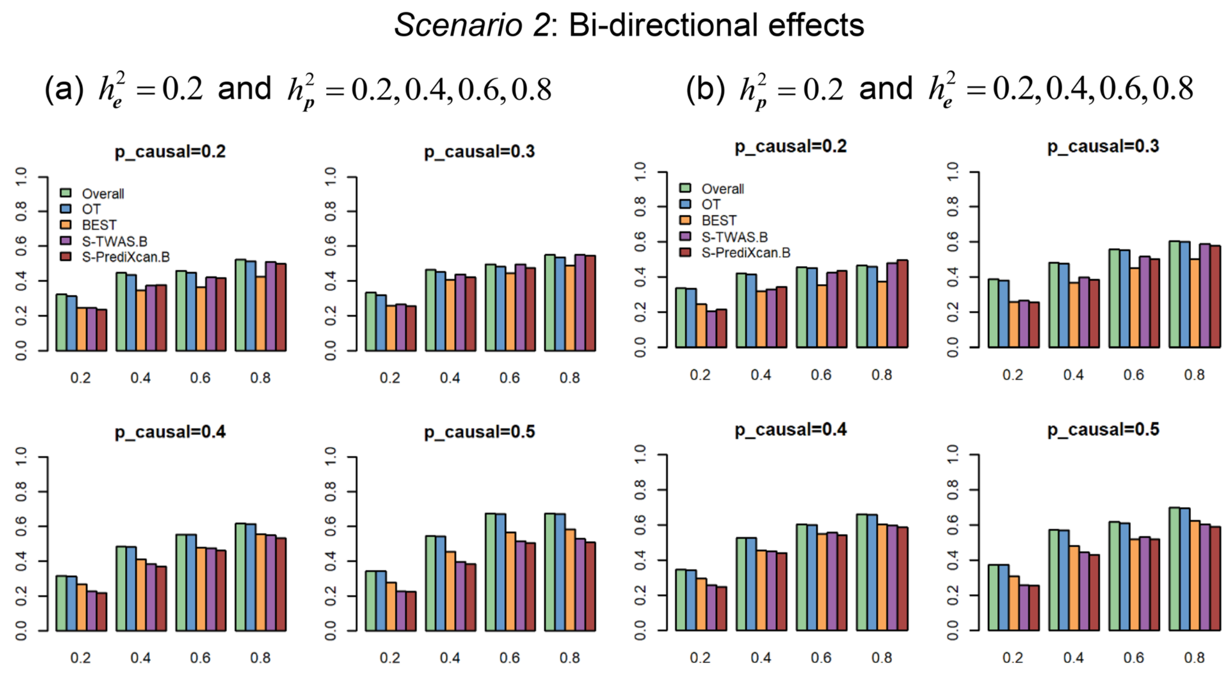

Figure 2. Power comparisons of gene-based association tests at 1.75×10^{-6} significance level under Bi-directional effects ($\beta_1 = \beta_2 = -\beta_3 = -\beta_4$) with $p_{causal} = (0.2, 0.3, 0.4, 0.5)$ based on gene *AGTRAP*. (**a**) Estimated power against phenotypic heritability h_p^2 with expression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

42 genetic variants at *C3orf22*. We vary the proportion of causal variants with values $p_{causal} = (0.2, 0.3, 0.4, 0.5)$ for *AGTRAP* and $p_{causal} = (0.1, 0.2, 0.3, 0.4)$ for *C3orf22*. We also consider two different directions of genetic effects: $\beta_1 = \cdots = \beta_K$ (Scenario 1: Uni-directional effects) and $\beta_1 = \cdots = \beta_{K/2} = -\beta_{K/2+1} = \cdots = -\beta_K$ (Scenario 2: Bi-directional effects). For each simulation scenario, we vary the proportion of gene expression heritability

and the phenotypic heritability with different values of h_e^2 and h_p^2. We consider the sample size to be 2000 (unless it is specified) and the power is calculated as the proportion of 1000 replications with p-value $< 1.75 \times 10^{-6}$.

Figure 1 (Supplementary Fig. S4) show the power comparisons based on gene *AGTRAP* (and *C3orf22*) with $K = 4$ under the Uni-directional effects ($\beta_1 = \beta_2 = \beta_3 = \beta_4$) with different p_{causal}. We consider two settings here. First, we vary phenotypic heritability h_p^2 with a fixed expression heritability $h_e^2 = 0.2$ (Fig. 1a and Supplementary Fig. S4a). Second, we vary the expression heritability h_e^2 with a fixed phenotypic heritability $h_p^2 = 0.2$ (Fig. 1b and Supplementary Fig. S4b). Figure 2 (Supplementary Fig. S5) shows power comparisons based on gene *AGTRAP* (and *C3orf22*) under the Bi-directional effects ($\beta_1 = \beta_2 = -\beta_3 = -\beta_4$) with different p_{causal} for $K = 4$. We also consider two simulation settings, power against the phenotypic heritability h_p^2 with a fixed expression heritability $h_e^2 = 0.2$ and power against the expression heritability h_e^2 with a fixed phenotypic heritability $h_p^2 = 0.2$. The pattern of the power in Fig. 2 (Supplementary Fig. S5) is similar to what we observe in Fig. 1 (Supplementary Fig. S4). These figures show that (1) Overall and OT perform uniformly better than BEST, S-TWAS.B, and S-PrediXcan.B. We can see that Overall and OT boost power significantly due to integrating association evidence by different traditional tests and multiple eQTL—derived weights. Overall is slightly more powerful than OT in all of the scenarios. (2) Among BEST, S-TWAS.B, and S-PrediXcan.B, BEST is more powerful than S-TWAS.B and S-PrediXcan.B in all of the scenarios for gene *C3orf22*; For gene *AGTRAP*, S-TWAS.B and S-PrediXcan.B perform better than BEST when the proportion of causal variants in a gene is small ($p_{causal} = (0.2, 0.3)$); otherwise, BEST performs better than S-TWAS.B and S-PrediXcan.B.

To evaluate if Overall and OT that integrate different types of association tests and multiple eQTL—derived weights are robust for more eQTL studies, we also consider 20 ($K = 20$) eQTL—derived weights under Uni-directional effect and Bi-directional effect models on gene *C3orf22* with settings similar to the settings in Supplementary Figs. S4 and S5. After integrating $L = 3(K + 1) = 63$ traditional gene-based association tests, we observe that the patterns of the power for $K = 20$ are similar to that in Supplementary Figs. S4 and S5 with $K = 4$, and the power gain of Overall and OT is higher than that of the tests only consider one eQTL—derived weight, such as BEST, S-PrediXcan.B, and S-TWAS.B (Supplementary Fig. S6).

Furthermore, we consider simulation settings with noise to the eQTL. We consider simulation settings by adding less noise to the eQTL from the most relevant tissues and more noise to those from the less relevant tissues. For the Uni-direction scenario, we consider the first study being the most relevant tissue, where $\beta_1 = \beta_0 + N\left(0, 0.1h_p^2\right)$ and $\beta_2 = \beta_3 = \beta_4 = \beta_0 + N\left(0, 0.5h_p^2\right)$; $\beta_0 = \sqrt{h_p^2 / K}$ depends on the phenotypic heritability h_p^2. For the Bi-direction scenario, we consider the first and third studies being the most relevant tissues that have opposite effect directions, where $\beta_1 = -\beta_0 + N\left(0, 0.1h_p^2\right)$, $\beta_3 = \beta_0 + N\left(0, 0.1h_p^2\right)$, and $\beta_2 = -\beta_0 + N\left(0, 0.5h_p^2\right)$, $\beta_4 = \beta_0 + N\left(0, 0.5h_p^2\right)$. Other parameter settings are the same as these in Supplementary Figs. S4 and S5. The power comparison results are shown in Supplementary Figs. S7 and S8. From these figures, we find that the patterns of the power in Supplementary Figs. S7 and S8 are very similar to those in Supplementary Figs. S4 and S5.

In all of the previous power comparisons, we use a sample size of 2000. We also consider simulation settings as those in Supplementary Figs. S7 and S8, but with a large sample size of 100,000. Supplementary Figure S9 shows the results of power comparisons. We can see from this figure, all powers are increased with this larger sample size, but the patterns of the power are very similar to those in Supplementary Figs. S7 and S8.

To remove noise in LD matrix computed from a reference sample, we shrink the observed LD matrix toward an identity matrix with the shrinkage parameter estimated by maximum likelihood. To evaluate how well this regulation process performs, we compare the powers of three traditional gene-based association tests with and without eQTL—derived weights, OT, and Overall based on corrected and uncorrected LD structure. We use the same simulation settings as those in Supplementary Figs. S7 and S8. Supplementary Figure S10 shows the power comparison results based on gene *C3orf22* under Uni-directional effects and Bi-directional effects with noise to eQTL. We can see that the powers of these tests based on corrected LD structure perform better than those based on uncorrected LD structure in most of the settings.

Real data analysis

To evaluate the performance of our proposed method, we apply Overall, OT, the three traditional tests with and without eQTL—derived weights, S-PrediXcan, and S-TWAS to the GWAS summary statistics data sets used in Zhang et al.[15]: two SCZ GWAS summary data sets and two lipid GWAS summary data sets. We estimate the LD between genetic variants using the 1000 Genomes Project reference panel[16], and obtain the corrected matrix of Z-score after the regularization procedure. We consider four eQTL—derived weights estimated by the BLUP method using the resources listed in Table 1 (NTR, YFS, METSIM, CMC).

Application to the SCZ GWAS summary data.

We consider two SCZ GWAS summary data sets, SCZ1 and SCZ2, which can be downloaded from the Psychiatric Genomics Consortium website (https://www.med.unc.edu/pgc/results-and-downloads/)[33]. SCZ1 is a meta-analysis of SCZ GWAS data set with 13,833 cases and 18,310 controls. SCZ2 is a more recent and larger SCZ GWAS summary data set with 36,989 cases and 113,075 controls for partial validation[34]. In our real data analysis, we define a gene to include all of the SNPs from 20 kb upstream to 20 kb downstream of the gene and test the association between each gene and the trait. We consider all genes according to the GENCODE version 35 (GRCh37) human comprehensive gene annotation list which can be downloaded from the GENCODE website (https://www.gencodegenes.org/human/release_35lift37.html).

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

Table 3. The numbers of genes identified by each method for the two SCZ data sets. The subscript denotes BT, SKAT, and SKATO using eQTL—derived weights; CMC, METSIM, NTR, and YFS indicate the resources to obtain the eQTL—derived weights. 0 indicates the methods without using any weights. SCZ1 indicates the number of genes identified by each method for SCZ1 data; SCZ2 indicates the number of genes identified by each method for SCZ2 data; $SCZ_{overall}$ indicates the number of overlapping genes identified by both SCZ1 and SCZ2 data sets; $GWAS_{SCZ1}$ and $GWAS_{SCZ2}$ indicate the numbers of genome-wide significant genes that are reported in the GWAS catalog and are also identified by each method for SCZ1 and SCZ2, respectively.

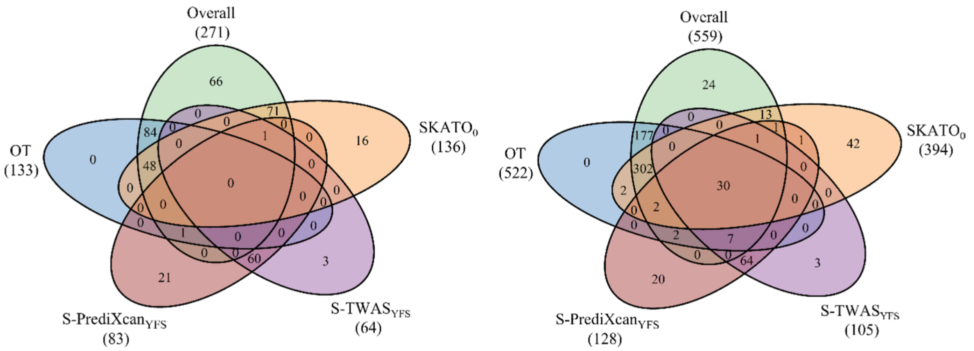

Figure 3. Venn diagram of the number of genes identified by Overall, OT, and $SKATO_0$, S-$PrediXcan_{YFS}$, and S-$TWAS_{YFS}$ for SCZ1 data (left) and SCZ2 data (right). The number below each method indicates the total number of significant genes identified by the corresponding method.

To make fair comparisons among all these weighted tests, the genetic variants are removed if there is at least one weight missing in the four eQTL—derived weights. After pruning, there are 26,575 genes in SCZ1 and 17,823 genes in SCZ2 left in our final analyses. Therefore, the Bonferroni corrected significance level for gene-based association analysis is defined as 0.05 divided by the number of genes. First, we apply BT, SKAT, and SKATO with and without an eQTL—derived weight, OT, Overall, S-PrediXcan, and S-TWAS to the SCZ1 and SCZ2 data sets. Table 3 (SCZ1 and SCZ2) shows the number of genes identified by each method for the SCZ data sets, respectively. As we can see in Table 3, Overall identifies more genes than all of the other methods for two SCZ

Gene	Data	Overall	References
RAI1	SCZ1	2.63E−31	Pardiñas et al.[35]
SLC7A6	SCZ1	2.17E−15	Ikeda et al.[36]; Li et al.[37]
AP001931.2	SCZ1	1.27E−13	Schizophrenia Working Group of the Psychiatric Genomics Consortium[34]; Goes et al.[38]; Ikeda et al.[36]; Li et al.[37]; Lam et al.[39]; Periyasamy et al.[40]; Lee et al.[41]; The Autism Spectrum Disorders Working Group of the Psychiatric Genomics Consortium[42]; Pardiñas et al.[35]
MARK2	SCZ1	2.64E−07	Goes et al.[38]
GULOP	SCZ1	1.24E−07	Pardiñas et al.[35]; Ikeda et al.[36]; Li et al.[37]; Goes et al.[38]; Lam et al.[43]
ZBED4	SCZ1	9.02E−07	Goes et al.[38]
RAB11FIP5	SCZ2	1.05E−06	Goes et al.[38]; Lam et al.[43]
AL669918.1	SCZ2	2.03E−06	Goes et al.[38]
YPEL1	SCZ2	2.80E−06	Goes et al.[38]
LINC00606	SCZ2	2.57E−06	Goes et al.[38]
ERLIN1	SCZ2	2.34E−06	Goes et al.[38]
AC024597.1	SCZ2	2.56E−06	Lam et al.[39]

Table 4. Genes identified only by Overall based on the two SCZ data sets that are reported in the GWAS catalog.

GWAS summary data sets. Among the three types of gene-based association tests, BT, SKAT, and SKATO, with or without different eQTL—derived weights, $SKATO_0$ identifies the largest number of genes. $S\text{-}TWAS_{YFS}$ and $S\text{-}PrediXcan_{YFS}$ identify the largest number of genes compared with S-TWAS and PrediXcan based on the other three eQTL—derived weights, respectively. Therefore, in Fig. 3, we only show the number of genes identified by Overall, OT, $SKATO_0$, $S\text{-}PrediXcan_{YFS}$, and $S\text{-}TWAS_{YFS}$. The number below each method indicates the total number of genes identified by the corresponding method. From Fig. 3, we can see that Overall identifies all of the genes identified by OT for SCZ1; for SCZ2, there are two genes identified by OT but failed to be identified by Overall; there are 66 and 24 genes identified only by Overall for SCZ1 data and SCZ2, respectively.

We further investigate the 90 genes identified only by Overall for the SCZ data sets by searching the GWAS catalog (https://www.ebi.ac.uk/gwas/). Among the 66 genes for the SCZ1 data set, there are six genes reported in the GWAS catalog; among the 24 genes for the SCZ2 data set, there are six genes reported in the GWAS catalog (Table 4). We also use these two SCZ GWAS data sets for partial validation. Table 3 shows that there are 45 overlapping genes identified by Overall using SCZ1 and SCZ2 data sets and only 17 overlapping genes identified by OT using both SCZ1 and SCZ2 data sets. Furthermore, we search for genome-wide significant SNPs ($p < 5 \times 10^{-8}$) from the two SCZ GWAS summary data sets and consider the genes covering at least one genome-wide significant SNP from 20 kb upstream to 20 kb downstream of the gene. There are 63 genome-wide significant genes for SCZ1, and 2422 genome-wide significant genes in SCZ2. Table 3 ($GWAS_{SCZ1}$ and $GWAS_{SCZ2}$) summarizes the numbers of genome-wide significant genes that are identified by each method for the two SCZ data sets. Among the 63 genome-wide significant genes for the SCZ1 data set, Overall identifies the largest number of genes, followed by $SKAT_0$ and $SKATO_0$; OT, $S\text{-}PrediXcan_{NTR}$ and $S\text{-}TWAS_{NTR}$ only identify 6 genes. Meanwhile, among 2422 genome-wide significant genes for SCZ2, Overall identifies 167 genes; OT identifies 166 genes; SKATO and $SKATO_0$ identify 153 genes; $S\text{-}TWAS_{YFS}$ and $S\text{-}PrediXcan_{YFS}$ only identify 58 and 72 genes respectively.

Application to the lipids GWAS summary data. We consider two lipids GWAS summary data sets, HDL1 and HDL2, which can be downloaded at the Center for Statistical Genetics (CSG) at the University of Michigan. HDL1 is a meta-analysis of HDL GWAS data set with about 100,000 samples downloaded at the website (http://csg.sph.umich.edu/willer/public/lipids2010/)[44]. HDL2 is the follow-up data with about 189,000 samples for partial validation downloaded at the Global Lipids Genetics Consortium (http://csg.sph.umich.edu/willer/public/lipids2013/)[45]. We perform the same analysis as we did in the previous section for the two SCZ GWAS summary data sets. After pruning and removing the genetic variants with missing weights, there are 17,389 genes in HDL1 and 16,917 genes in HDL2. Table 5 (HDL1 and HDL2) shows the number of genes identified by each method for the two lipids data sets, respectively. As we can see from Table 5, among the three traditional gene-based association tests with and without eQTL—derived weights, $SKATO_0$ and BT_0 identify the largest number of genes in HDL1 and HDL2, respectively; Among the four S-PrediXcan tests, $S\text{-}PrediXcan_{YFS}$ and $S\text{-}PrediXcan_{CMC}$ identify the largest number of genes in HDL1 and HDL2, respectively; for the four S-TWAS tests, $S\text{-}TWAS_{YFS}$ and $S\text{-}TWAS_{CMC}$ identify the largest number of genes in HDL1 and HDL2, respectively. For the HDL1 data set, Overall identifies the largest number of genes (249), followed by OT that identifies 233 genes; for the HDL2 data set, BT_0 identifies the largest number of genes (836), followed by Overall and OT, where Overall identifies 765 genes and OT identifies 688 genes. In Fig. 4, we compare genes identified by $SKATO_0$, $S\text{-}PrediXcan_{YFS}$, and $S\text{-}TWAS_{YFS}$, along with Overall and OT for the HDL1 data set and genes identified by BT_0, $S\text{-}PrediXcan_{CMC}$, $S\text{-}TWAS_{CMC}$, Overall, and OT for the HDL2 data set. Again, we observe that Overall identifies the largest number of genes for the HDL1 data set and the second most for the HDL2 data set; all genes identified by OT are also identified by Overall; 82 and 24 genes are identified only by Overall and OT for the HDL1 and HDL2 data sets, respectively; there are 13 and 6 genes only identified by Overall for the HDL1 and HDL2 data sets, respectively. We search the GWAS catalog (https://www.ebi.ac.uk/gwas/). Table 6 shows that five out

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

Table 5. The number of genes identified by each method for the two lipids data sets. The subscript denotes BT, SKAT, and SKATO using eQTL—derived weights; CMC, METSIM, NTR, and YFS indicate the resources to obtain the eQTL—derived weights. 0 indicates the methods without using any weights. HDL1 indicates the number of genes identified by each method for HDL1 data; HDL2 indicates the number of genes identified by each method for HDL2 data; $HDL_{overall}$ indicates the number of overlapping genes identified by both HDL1 and HDL2 data sets; $GWAS_{HDL1}$ and $GWAS_{HDL2}$ indicate the numbers of genome-wide significant genes that are reported in the GWAS catalog and are also identified by each method for HDL1 and HDL2, respectively.

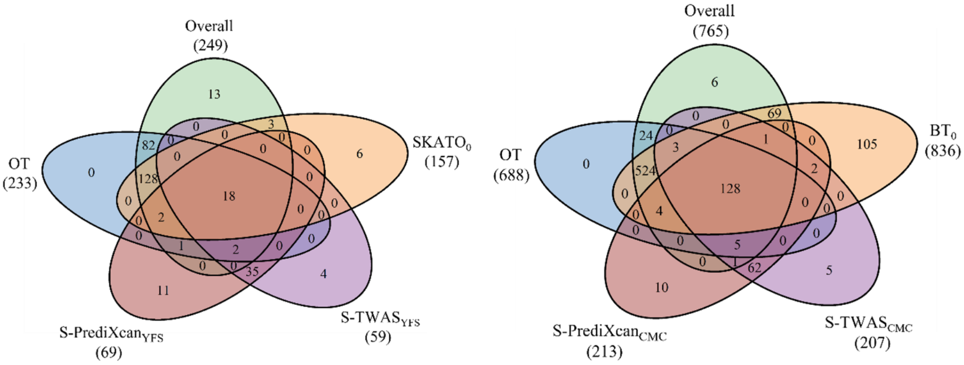

Figure 4. Venn diagram of the number of genes identified by Overall, OT, $SKATO_0$, S-$PrediXcan_{YFS}$, and S-$TWAS_{YFS}$ for HDL1 data (left) and Overall, OT, BT_0, S-$PrediXcan_{CMC}$, and S-$TWAS_{CMC}$ for HDL2 data (right). The number below each method indicates the total number of significant genes identified by the corresponding method.

of 13 genes identified only by Overall based on HDL1 data have been reported, and one out of 6 genes has been reported on HDL2 data in the GWAS catalog. We also use these two HDL GWAS data sets for partial validation by looking for the number of overlapping genes identified by both of the data sets (Table 5, $HDL_{overlap}$). There are 177 overlapping genes identified by Overall for both SCZ1 and SCZ2 data sets and 167 overlapping genes identified by OT for both SCZ1 and SCZ2 data sets.

Gene	Data	Overall	References
AP002954.1	HDL1	2.27E−11	Emilsson et al.[46]
EDC4	HDL1	1.65E−11	Lettre et al.[47], Kilpeläinen et al.[48], Wojcik et al.[49]
PACSIN1	HDL1	2.24E−06	Liu et al.[50]
AFF1	HDL1	2.10E−06	Spracklen et al.[51], De Vries et al.[52], Hoffmann et al.[53], Ripatti et al.[54], Richardson et al.[55]
AC106779.1	HDL1	2.85E−06	Noordam et al.[56]
NHLRC2	HDL2	1.98E−06	Hoffmann et al.[53], Richardson et al.[55], Klarin et al.[57], Qi et al.[58], Klimentidis et al.[59], Liu et al.[60]

Table 6. Genes identified only by Overall based on the two lipids data sets that are reported in the GWAS catalog.

Same as the analyses for the SCZ GWAS summary data sets, we search for genome-wide significant SNPs ($p < 5 \times 10^{-8}$) from the two lipids GWAS summary statistics. There are 1911 genome-wide significant genes for HDL1 and 2682 genome-wide significant genes for HDL2. Table 5 (GWAS$_{\text{HDL1}}$ and GWAS$_{\text{HDL2}}$) summarizes the numbers of genome-wide significant genes that are identified by each method for the two lipids data sets. Among the 1911 genome-wide significant genes for the HDL1 data set, Overall identifies the largest number of genes (122), followed by OT (120), then SKAT$_0$ (104); S-TWAS$_{\text{YFS}}$ only identifies 29 genes and S-PrediXcan$_{\text{YFS}}$ identifies 31 genes. Meanwhile, among 2682 genome-wide significant genes for HDL2, Overall identifies the largest number of genes (192); OT and SKATO$_0$ identify 190 genes; S-TWAS$_{\text{METSIM}}$ and S-PrediXcan$_{\text{METSIM}}$ identify 112 and 118 genes. respectively.

Discussions

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

There are some advantages of our proposed method. First, Overall adaptively aggregates information from multiple gene-based association tests. Most combination tests (i.e., Fisher's combination test[61]) assume that the p-values should be calculated from independent tests. To combine information from highly correlated gene-based association tests, Overall utilizes the extended Simes procedure[5,22]. It is shown that this procedure to combine multiple tests is stable and effective regardless of whether the tests are highly correlated[24,62]. Second, Overall is more powerful than the traditional gene-based association tests, some popular transcriptome association tests (i.e., S-PrediXcan[29] and S-TWAS[12]), and other eQTL weighted combination tests (i.e., ominous test[15]). By aggregating information from different tests and incorporating multiple eQTL—derived weights, Overall can achieve a higher statistical power under a variety of situation settings. Meanwhile, our simulation studies and real data analyses show that the extended Simes procedure is more powerful than the Cauchy combination method, especially if the proportion of causal variants in a gene is small. Third, the p-values of Overall can be analytically computed without using permutations, therefore, Overall is computationally efficient. Finally, using the regularization procedure to correct the estimated LD can reduce the potential statistical noise in the LD estimation if LD is estimated using a reference panel with small sample size. In addition, Overall can be easily applied to genetic association studies with either individual-level data or GWAS summary statistics.

In this paper, we combine three types of traditional gene-based association tests (BT, SKAT, SKATO). However, the combination procedure used in the paper is very general. Other more powerful gene-based association tests can also be combined using the same approach, such as some state-of-the-art methods (i.e., S-TWAS[12], E-MAGMA[63], and SMR[64]).

In this current study, we utilize the weights derived from four single tissue gene expression studies (CMC, METSIM, NTR, YFS). Although the extended Simes procedure in Overall allows us to employ more eQTL—derived weights from a number of studies (i.e., GTEx gene expression version 8[65] et al.), there is a possibility that the noise can be increased with the increment in the number of unrelated studies. Therefore, the power of the combination tests (i.e., Overall and OT) might be attenuated. Thus, to obtain the most robust identification of phenotypic associated genes in a real data analysis with the Overall method, we suggest incorporating eQTL datasets from the most relevant tissues to the phenotype. The last but the most important thing is that population stratification can be confounded association results[66,67]. Systematic minor allele frequency difference between transcriptomic studies of different cohorts and no matching between the estimated LD structure of Genomes Project with that in the study may increase the chances of false positive findings. Therefore, we need to eliminate false positive findings possibly caused by population stratification[68,69]. When applying the Overall method, the population of GWAS summary dataset, external reference panel (i.e., 1000 Genomes Project) used to estimate LD structure, and eQTL—derived weights should be consistent.

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

Data availability

The data that support the findings of this study are publically available and the links to the data are provided in the article.

Received: 10 June 2021; Accepted: 11 February 2022
Published online: 03 March 2022

References

1. Fine, R. S., Pers, T. H., Amariuta, T., Raychaudhuri, S. & Hirschhorn, J. N. Benchmarker: An unbiased, association-data-driven strategy to evaluate gene prioritization algorithms. *Am. J. Hum. Genet.* **104**, 1025–1039 (2019).
2. Li, R. *et al.* A regression framework to uncover pleiotropy in large-scale electronic health record data. *J. Am. Med. Inform. Assoc.* **26**, 1083–1090 (2019).
3. Hebbring, S. J. The challenges, advantages and future of phenome-wide association studies. *Immunology* **141**, 157–165 (2014).
4. Kraft, P., Zeggini, E. & Ioannidis, J. P. Replication in genome-wide association studies. *Stat. Sci. Rev. J. Inst. Math. Stat.* **24**, 561 (2009).
5. Li, M.-X., Gui, H.-S., Kwan, J. S. & Sham, P. C. GATES: A rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet.* **88**, 283–293 (2011).
6. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
7. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
8. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
9. Svishcheva, G. R., Belonogova, N. M., Zorkoltseva, I. V., Kirichenko, A. V. & Axenovich, T. I. Gene-based association tests using GWAS summary statistics. *Bioinformatics* **35**, 3701–3708 (2019).
10. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).
11. Conneely, K. N. & Boehnke, M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am. J. Hum. Genet.* **81**, 1158–1168 (2007).
12. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
13. Kwak, I.-Y. & Pan, W. Adaptive gene-and pathway-trait association testing with GWAS summary statistics. *Bioinformatics* **32**, 1178–1184 (2016).
14. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
15. Zhang, J., Xie, S., Gonzales, S., Liu, J. & Wang, X. A fast and powerful eQTL weighted method to detect genes associated with complex trait using GWAS summary data. *Genet. Epidemiol.* **44**, 550–563 (2020).
16. Consortium, G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
17. Deng, Y. & Pan, W. Improved use of small reference panels for conditional and joint analysis with GWAS summary statistics. *Genetics* **209**, 401–408 (2018).
18. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
19. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
20. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091 (2015).
21. Xu, Z., Wu, C., Wei, P. & Pan, W. A powerful framework for integrating eQTL and GWAS summary data. *Genetics* **207**, 893–902 (2017).
22. Van der Sluis, S., Posthuma, D. & Dolan, C. V. TATES: Efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet.* **9**, e1003235 (2013).
23. Dutta, D. *et al.* A powerful subset-based method identifies gene set associations and improves interpretation in UK Biobank. *Am. J. Hum. Genet.* **108**, 669–681 (2021).
24. Wu, C. Multi-trait genome-wide analyses of the brain imaging phenotypes in UK Biobank. *Genetics* **215**, 947–958. https://doi.org/10.1534/genetics.120.303242 (2020).
25. Yang, Y., Basu, S., Mirabello, L., Spector, L. & Zhang, L. A Bayesian gene-based genome-wide association study analysis of osteosarcoma trio data using a hierarchically structured prior. *Cancer Inform.* **17**, 1176935118775103 (2018).
26. Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* **93**, 42–53 (2013).
27. Hogg, R. V., Tanis, E. A. & Zimmerman, D. L. *Probability and Statistical Inference*. vol. 993 (Macmillan New York, 1977).
28. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).
29. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1–20 (2018).
30. Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **115**, 393–402 (2020).
31. Rosseel, Y. Lavaan: An R package for structural equation modeling and more. Version 0.5-12 (BETA). *J. Stat. Softw.* **48**, 1–36 (2012).
32. Nagpal, S. *et al.* TIGAR: An improved Bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *Am. J. Hum. Genet.* **105**, 258–266 (2019).
33. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150 (2013).
34. Consortium, S. W. G. O. T. P. G. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
35. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
36. Ikeda, M. *et al.* Genome-wide association study detected novel susceptibility genes for schizophrenia and shared trans-populations/diseases genetic effect. *Schizophr. Bull.* **45**, 824–834 (2019).
37. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576 (2017).

38. Goes, F. S. *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **168**, 649–659 (2015).
39. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* **51**, 1670–1678 (2019).
40. Periyasamy, S. *et al.* Association of schizophrenia risk with disordered niacin metabolism in an Indian genome-wide association study. *JAMA Psychiat.* **76**, 1026–1034 (2019).
41. Lee, P. H. *et al.* Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469–1482 (2019).
42. The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Mol. Autism* **8**, 1–17 (2017).
43. Lam, M. *et al.* Pleiotropic meta-analysis of cognition, education, and schizophrenia differentiates roles of early neurodevelopmental and adult synaptic pathways. *Am. J. Hum. Genet.* **105**, 334–350 (2019).
44. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
45. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274 (2013).
46. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).
47. Lettre, G. *et al.* Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS Genet.* **7**, e1001300 (2011).
48. Kilpeläinen, T. O. *et al.* Multi-ancestry study of blood lipid levels identifies four loci interacting with physical activity. *Nat. Commun.* **10**, 1–11 (2019).
49. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
50. Liu, H. *et al.* Heritability and genome-wide association study of plasma cholesterol in Chinese adult twins. *Front. Endocrinol.* **9**, 677 (2018).
51. Spracklen, C. N. *et al.* Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels. *Hum. Mol. Genet.* **26**, 1770–1784 (2017).
52. De Vries, P. S. *et al.* Multiancestry genome-wide association study of lipid levels incorporating gene-alcohol interactions. *Am. J. Epidemiol.* **188**, 1033–1054 (2019).
53. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).
54. Ripatti, P. *et al.* Polygenic hyperlipidemias and coronary artery disease risk. *Circ. Genom. Precis. Med.* **13**, e002725 (2020).
55. Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS Med.* **17**, e1003062 (2020).
56. Noordam, R. *et al.* Multi-ancestry sleep-by-SNP interaction analysis in 126,926 individuals reveals lipid loci stratified by sleep duration. *Nat. Commun.* **10**, 1–13 (2019).
57. Klarin, D. *et al.* Genetics of blood lipids among~ 300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
58. Qi, G. & Chatterjee, N. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. *PLoS Genet.* **14**, e1007549 (2018).
59. Klimentidis, Y. C. *et al.* Phenotypic and genetic characterization of lower LDL cholesterol and increased type 2 diabetes risk in the UK Biobank. *Diabetes* **69**, 2194–2205 (2020).
60. Liu, D. J. *et al.* Exome-wide association study of plasma lipids in > 300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
61. Curtis, D., Vine, A. E. & Knight, J. A simple method for assessing the strength of evidence for association at the level of the whole gene. *Adv. Appl. Bioinform. Chem. AABC* **1**, 115 (2008).
62. Wang, M. *et al.* COMBAT: A combined association test for genes using summary statistics. *Genetics* **207**, 883–891 (2017).
63. Gerring, Z. F., Mina-Vargas, A., Gamazon, E. R. & Derks, E. M. E-MAGMA: An eQTL-informed method to identify risk genes using genome-wide association study summary statistics. *Bioinformatics* **37**, 2245–2249 (2021).
64. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
65. Feng, H. *et al.* Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. *PLoS Genet.* **17**, e1008973 (2021).
66. Sha, Q., Wang, X., Wang, X. & Zhang, S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genet. Epidemiol.* **36**, 561–571 (2012).
67. Zhu, H., Zhang, S. & Sha, Q. A novel method to test associations between a weighted combination of phenotypes and genetic variants. *PLoS One* **13**, e0190788 (2018).
68. Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
69. Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388–393 (2004).

Acknowledgements

Xuexia Wang was supported by the University of North Texas Foundation which was contributed by Dr. Linda Truitt Creagh. High-Performance Computing Shared Facility (Superior) at Michigan Technological University was used in obtaining results presented in this publication. Some parts of this work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges-2 system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

Author contributions

Formal analysis: X.C.; methodology: X.C., S.Z., X.W., and Q.S.; data curation: X.C. and X.W.; visualization: X.C.; writing original draft: X.C., X.W., and Q.S.; writing review and editing: X.C., S.Z., X.W., and Q.S.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-07465-0.

Correspondence and requests for materials should be addressed to Q.S.

Reprints and permissions information is available at www.nature.com/reprints.